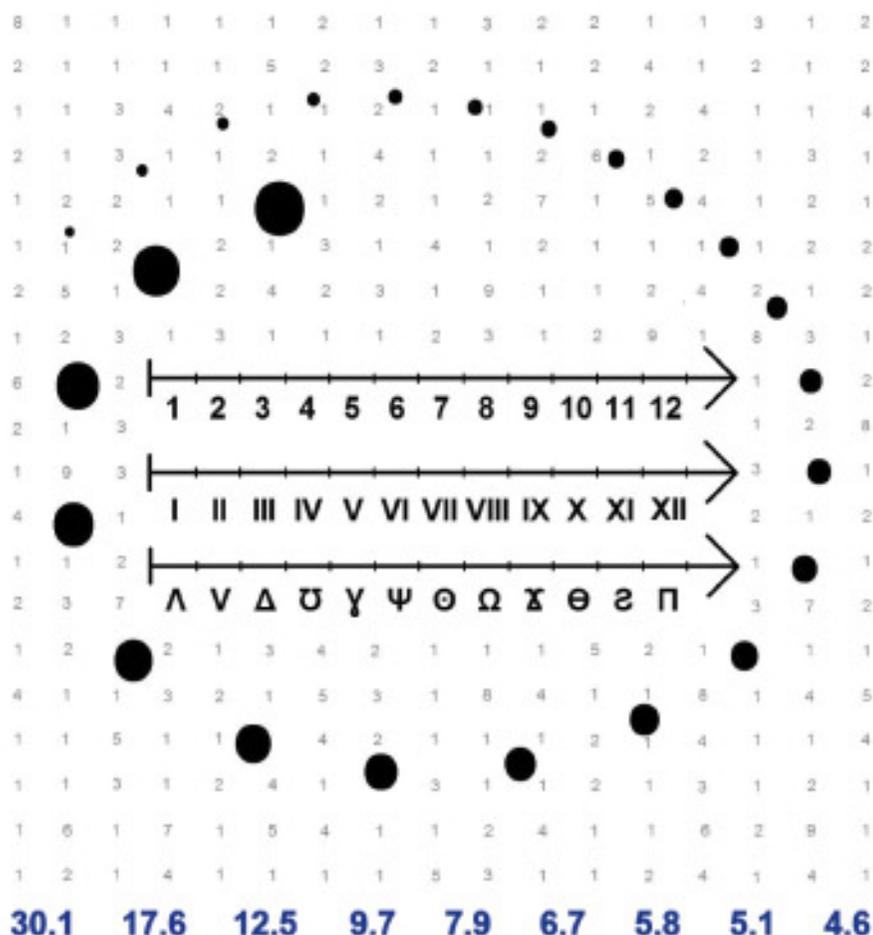


# Benford's Law

Theory, the General Law of Relative Quantities,  
and Forensic Fraud Detection Applications

Alex Ely Kossovsky



# **Benford's Law**

**Theory, the General Law of Relative Quantities,  
and Forensic Fraud Detection Applications**

**This page intentionally left blank**

# **Benford's Law**

**Theory, the General Law of Relative Quantities,  
and Forensic Fraud Detection Applications**

**Alex Ely Kossovsky**

*The City University of New York, USA*

 **World Scientific**

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

*Published by*

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

**Library of Congress Cataloging-in-Publication Data**

Kossovsky, Alex Ely.

Benford's law : theory, the general law of relative quantities, and forensic fraud detection applications / Alex Ely Kossovsky.

pages cm

Includes bibliographical references and index.

ISBN 978-9814583688 (hardcover : alk. paper)

1. Fraud investigation--Statistical methods. 2. Fraud--Statistical methods. 3. Distribution (Probability theory) 4. Forensic statistics. 5. Forensic sciences--Statistical methods. I. Title.

HV8079.F7K67 2014

363.25'963--dc23

2014024759

**British Library Cataloguing-in-Publication Data**

A catalogue record for this book is available from the British Library.

Copyright © 2015 by World Scientific Publishing Co. Pte. Ltd.

*All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the publisher.*

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

Typeset by Stallion Press

Email: [enquiries@stallionpress.com](mailto:enquiries@stallionpress.com)

Printed in Singapore

## BENFORD'S LAW

---

---

*The mathematicians believe it [the Normal Distribution] to be a physical fact while the scientists believe it to be a mathematical law*

Poincaré (1854–1912)

Benford's Law has several names and many discoverers, inspired by a bizarre array of topics including old logarithm and trigonometry reference books, broken calculator keys, burned-out vacuum tubes, river lengths, and postal addresses. Each generation seems to find a new version of this digital law, with new applications in mathematics, statistics, number theory, quality control, forensic accounting, and even metaphysics. Just like the fable about the blind men encountering an elephant and describing it by touch alone — an elephant is like a snake, a tree, a fan, a rope, a spear — quantitative investigators of all stripes often fail to see the consequences of Benford's Law outside their own field of expertise. What is lacking is a treatment of the entire subject, where the forest can be seen fully in all its glory and beauty, instead of merely individual and unconnected trees.

Now that lack has been remedied by the masterful manuscript by Alex Ely Kossovsky. All of Benford's Law is examined over the vast range of subjects described above. There seems to be no obscure journal he has not prowled, and no practitioner he has not interviewed. But this is no mere compilation. Kossovsky has found several new insights into Benford's Law:

- Its application not only to mixtures of distributions, but what he calls distribution chains: a distribution whose parameters are subjects of a second distribution, which itself is governed by a third distribution, which is ruled by which is determined by a fourth, and so on. (Benford's Law as the ultimate Bayesian prior!)
- The Benford effect on digits after the first order is well known, with the Exponential-like distribution decaying into the Uniform for higher digital orders, but Kossovsky is the first to mathematically characterize this “drift”.

- Similarly, the need for several orders of magnitude range in the data in order to obtain Benford behavior is well-known, but Kossovsky is the first to show just how many orders of magnitude are exactly needed and how to go about measuring them correctly.
- Benford's Law has novel implications for determining optimal bin size.
- Benford's Law effects exist even for non-numerical data!

The book ends with some fascinating speculations on the almost axiomatic status of Benford's Law at the deepest levels of mathematics, closing with Frank Benford's own final words asserting that the phenomenon is truly physical, not numerical or digital. A parallel quote from Richard Feynman:

...we have seen that the complexities of things can so easily and dramatically escape the simplicity of the equations that describe them... Yet, all of these are really in the equations; we just haven't found a way to get them out... The next great awakening of human intellect may well produce a method of understanding the qualitative content of equations. Today we cannot. (*Lectures in Physics II*, 41–42).

This book in a sense is part of that “next great awakening” Feynman describes, which today we would recognize as the Langlands conjecture regarding the unity of all mathematics.

Forty-five years ago, I was stunned and skeptical when my teacher, John Tukey, told us “about half of all numbers begin with 1 or 2”. Kossovsky's book will go a long way to decreasing the surprise and mistrust that is the initial reaction of most people when told of Benford's Law.

Prof. Edward Binkowski  
Department of Applied Mathematics and Statistics  
City University of New York, Hunter College

## FOREWORD

---

---

To avoid confusion, perhaps I should inform the reader right away that I'm not the Frank Benford for whom "Benford's Law" is named, but one of his grandchildren. I'm a professional applied mathematician and quite familiar with Benford's Law, but it isn't the principal focus of my research interests.

I suppose that my grandfather would not have been surprised if the curious mathematical phenomenon he had rediscovered soon disappeared again into obscurity. He would probably be astonished (and doubtlessly pleased) that interest in his "Law of Anomalous Numbers" has not only continued but has actually increased exponentially over the 75 years since the publication of his paper. To document this growth of interest in Benford's Law, I used data found in the Benford online bibliography to construct the following table showing the number of "Benford relevant publications" in each decade between 1940 and 2010:

Number of pubs by decade	
1940–49	5
1950–59	5
1960–69	18
1970–79	40
1980–89	72
1990–99	101
2000–09	416

Papers and other publications about Benford's Law may be sorted roughly into three classes: (1) theoretical papers (concerned with the mathematical foundations of Benford's Law, or giving extensions of the Law), (2) expositions aimed at "the intelligent layman," and (3) papers dealing with applications of the Law. The book you're holding includes material that falls into all three of these classes. It contains a clear exposition of Benford's Law, a considerable number of pioneering mathematical results, and a wealth of "real world" examples.

Alex Ely Kossovsky brings a lively intelligence, a unique imagination, and a novel perspective to the study of Benford's Law, and I am certain that this book will engender many new lines of research on my grandfather's Law of Anomalous Numbers.

Frank Benford  
BenfordAppliedMath.com  
Salem, Oregon

# INTRODUCTION

---

---

When I trace at my pleasure the windings to and fro of the heavenly bodies I no longer touch earth with my feet: I stand in the presence of Zeus himself and take my fill of ambrosia, food of the gods.

Ptolemy (90–168 AD) Alexandria

Five objectives regarding style, format, and content guided the writing of this book: (1) To gather and synthesize from as many articles and sources as time and energy permit, all the relevant results, explanations, causes, and aspects of this digital phenomenon, with the intent to cover the subject comprehensively. (2) To focus on the challenging issue of organizing the exposition of this complex subject, arranging it in a well-organized way. (3) To present the material in a clear and reader-friendly style, often giving concrete numerical examples on top of abstract formulas and generic claims. (4) To facilitate visualization and illustration via as many charts, diagrams, scatter plots, and tables, as needed. (5) To facilitate intuition via conceptual presentation of the topics whenever possible, including the use of analogies and parables.

Readers such as accountants, auditors, tax officials, financiers, investors, and the general public should get a good feel of the topic primarily by reading all the chapters in sections 1 and 2, without any need whatsoever to struggle with advanced mathematics or statistics. They could further advance their knowledge a great deal by reading Section 3 on Data Compliance Tests, although this section might prove a bit more challenging, in which case it is suggested to skip the last five chapters there, namely omitting chapters 43 to 47 which are relatively more involved. They could certainly omit entirely Sections 4, 5, 6, and 7 without losing too much in any practical way regarding applications, fraud detection, and forensic digital analysis.

Mathematicians and statisticians should get thoroughly familiar with Sections 1 and 3 before plunging into the more advanced Sections of 4 to 7. Section 2 is dedicated to Fraud Detection and can be omitted by such readers without any loss

in understanding the rest of the book. Yet, beyond the exception of Section 2, readers are strongly advised to strictly follow the order of the book, since organization was quite deliberate, and reading the book in the order it is presented strongly facilitates the understanding of the material.

Section 7 — the transition from the digital point of view to a quantitative one — is imbedded in a fictional short story about a number-less society, serving as the backdrop of the whole discussion, in order to focus the mind of the reader, and also because this narrative is in part actually how I myself arrived at these new quantitative findings by simply imagining such numerically-backward society and wondering how would they be able to perceive Benford's Law — in spite of the obvious lack of digits.

Readers who have carefully read Section 1 already and are impatient to go in order, could still plunge directly into Section 7 (omitting chapters 114, 118, and 119 there) without any dramatic loss in understanding, and should be able to comprehend the main motivation behind the digital-to-quantitative revolution in outlook, as well as the main results outlined there, although this is not really the ideal order.

Sections 5 and 6 can also be read after a careful reading of Section 1 for the most part with only partial difficulties, although this is not really the ideal order. Within Section 6 itself which contains a variety of sub-topics in Benford's Law, there is a certain essential dependency and order of most of the chapters there, therefore the reader would be advised in general to go in the order they are actually presented there.

Writing up this book and building knowledge was a four-pronged process, involving: (1) scribbling mathematical equations and expressions, and manipulating terms; (2) examining on the computer large real-life data sets of physical, scientific, and financial records, to learn about their more detailed aspects, and potentially to get hints at some more general patterns; (3) Monte Carlo computer simulations of the relevant abstract distributions and generic random processes; and (4) developing conceptual frameworks and abstract general principles, making 'human sense' out of it all. Without any rigid order for these four activities above, one such activity usually led to another, which then suggested another activity, which pointed to yet another activity, and so forth. Not a single contradiction had ever emerged, no matter what angle I viewed things through, no matter what approach I used, the same results were always obtained, and perfect harmony and consistency prevailed, as if a benevolent and mystical force was always hovering just above me, guiding and supervising me, ensuring that the whole complex and delicate edifice does not fall

into ruin and contradictions. Scientists and especially pure mathematicians often take this fact for granted. One can derive the area of a circle in so many different ways, directly and indirectly. One can even physically carve a large 10-foot-radius circle made of cardboard and attempt to measure its area manually in the approximate by cutting out with a large scissor 1-by-1 square feet, counting them, and adding up the leftover ‘corners’ and ‘margins’ as much as possible; and no matter how it is done, no matter what distinct vista or algorithm is applied, miraculously  $\pi r^2$  always pops out, and no contradiction ever arises in the form of a shocking different result! The supposed implication here is that the universe — or at least our little corner of it — is thoroughly rational (and that it is only our collective social, political and environmental behavior that is severely irrational).

Pure mathematicians of the old school might be dismayed and quite upset by the inclusion of a method that I may perhaps characterize as “Mathematical Empiricism”, which is highly computer-dependend, and is akin to physically cutting out pieces from that circular cardboard to ‘directly measure’ its area. A more modern example of this method is letting the computer generate hundreds of thousands of prime numbers, followed by a construction of a crude histogram showing how many primes exist per N numbers, which then appears to resemble a great deal the gently falling curve of  $1/\ln(N)$ , meaning that primes become rarer and more sparse as one considers bigger numbers, and which is suggestive of the asymptotic law of distribution of prime numbers. Yet, without this essential method I could not have gotten the results I gathered; very little progress would have been made; mathematics would have been practiced for its own sake for enjoyment with limited relevance to the real world out there, and this book would probably not have been written at all. Fortunately for me and for this whole project, I live in the age of the personal computer and the Internet, and so in 5 minutes I could potentially download from, say, the NASA website, data on 50,000 stars, giving me their mass, luminosity, and distance from the Solar system, without polishing any telescopes, without painstakingly watching the sky for hours and gathering data, and all to be digitally analyzed on my computer for compliance with the ubiquitous law of Benford in less than a minute!

More than 130 years after the publication of an obscure article by Newcomb in 1881, an article which was subsequently ignored and then totally forgotten, the law is finally being fully acknowledged and widely utilized. Nowadays most governments around the world routinely apply Benford’s Law to detect tax fraud committed by their citizens and corporations. Unfortunately, the tables cannot be

turned around; digits and Benford's Law cannot enter the political fray by enabling citizens to examine and detect the fraud committed by their governments in undermining true, direct, and participatory democracy, the reduction of complex public lives of nations requiring multifaceted decisions into simplistic Boolean choices or a handful of candidates, including the possibility that positions and policy are legally allowed to be switched after elections, rendering the democratic process meaningless and deceptive. Nonetheless, several concrete academic studies using Benford's Law to detect electoral fraud have been published in recent years by political analysts equipped with the latest forensic digital techniques, with better results obtained by way of the 2nd order digit distribution. Since population data itself obeys Benford's Law almost perfectly, so does any percentage breakdown of counties and electoral districts, implying that highly skewed and unequal digital configuration should be found in all honest vote counts, while the corrupt political functionary concocting the counts would naturally do it in a totally random manner with regards to digits, thus yielding digital equality in the approximate. Here Benford's Law is utilized in a concrete way as a check against electoral fraud — and by extension against concealed tyranny — directly entering the political fray.

The reader could download from **[www.ForensicBenford.com](http://www.ForensicBenford.com)** all the data sets relevant to this book, including those used for the 11 case studies. In addition, several user-friendly MS-Excel files containing macros (programs) are included performing digital forensic analyses with a single click on the keyboard. At times I'm available for seminars and presentations, and can be contacted at the two email addresses below. Readers are welcome to write comments; argue about mathematical and statistical ideas; ask about forensic digital techniques and fraud detection issues; ask for help in using these forensic digital macros; or inquire about more advanced programming codes tailor-made for certain data configurations and those performing more sophisticated digital and data analyses.

Alex Ely Kossovsky

ForensicBenford.com  
akossovsky@gmail.com  
akossovsky@yahoo.com

## ACKNOWLEDGMENT

---

---

It is with the most heartfelt gratitude that I thank Professor Edward Binkowski of the Applied Mathematics and Statistics Department at The City University of New York, Hunter College, who has supported and encouraged me throughout my long journey in attempting to better understand and write about Benford's Law. Binkowski is a disciple (literally in a biblical sense so to speak) of John Tukey from Princeton University — one of the most influential statisticians of the last generation and known for coining the terms “bit” and “software”. It was during Binkowski's renowned Data Analysis class in the 2004 spring semester that he taught us about the existence of the phenomenon. Binkowski, who gave a presentation on Benford's Law years ago and is familiar with every page of this book, is well-known for being an exceptionally well-read academic, and for his extraordinary solid knowledge of diverse disciplines besides statistics and mathematics, including physics, philosophy of science, and history of mathematics, to mention just a few.

I wish to express my gratitude and debt to the distinguished mathematician George Andrews of Pennsylvania State University, considered to be the world's leading expert in the theory of integer partitions; well-known also for his extensive and dedicated work regarding Srinivasa Ramanujan manuscripts. I thank him for his assistance in obtaining the closed form expression of the general law of relative quantities, as well as for several insightful suggestions regarding the presentation of the 7th section — the only part of the manuscript he became familiar with, and which he considers to constitute a compelling argument of why the digital could or should be turned into the quantitative. I also thank Andrews for his kind invitation to Penn State for a short visit, for his exceptionally warm hospitality, and for his general support and encouragement in publishing the book.

I warmly thank the applied mathematician Frank Benford — the grandson of the discoverer who inherited his name (and evidently some of his mathematical abilities) — for reviewing the first two ‘practical’ sections and sending many helpful comments. I also wish to thank him for providing me with the best available

photo from the family albums of ‘the real’ Frank Benford, as well as for his general support for this book project.

I thank Ted Hill — the renowned mathematician who endowed the phenomenon with much greater mathematical respectability and is widely known as an authority in the field — for several highly enlightening and thought-provoking discussions about the phenomenon during his February 2012 academic visit to the University of Costa Rica. His enthusiasm about the field is refreshing and infectious, and I was very glad that I was able to share with him the wonder of how this exact digital proportional signature mysteriously presents itself in the deterministic as well as in the random.

I thank Ralph Raimi — one of the first mathematicians who had faith in Benford and his findings — for his many comments, suggestions, and especially for his enthusiastic endorsement of that newly found Digital Development Pattern which (to my dismay) he termed ‘digital footprints’ — as those markings left on the ground by a wild tiger [or another digital animal] passing a territory. Binkowski’s term of ‘left-center-right shift in distribution’ seems more acceptable, albeit a bit long. I will always prefer mine.

I thank the mathematician Steven Miller for first having faith early on in my outlandish conjecture regarding chains of distributions, and secondly for working with his students in his 2007 class to establish a rigorous mathematical proof for some particular cases of those chains involving standard distributions. I am still hoping for another more daring and generic rigorous proof which would encompass the principle in its more general scope as warranted by conceptual reasoning and solid Monte Carlo computer simulation results. I also wish to thank Miller for his kind invitation to join the Benford’s Law Conference in Santa Fe, New Mexico in 2007, and for finding a comfortable spot in the busy lecture schedule at the last moment for my talk about the chains of distributions.

I warmly thank the physicist Oded Kafri [the author of a recent book on entropy and thermodynamics] for spending considerable time with me in late 2008, patiently discussing his balls & boxes model, and his unique vista of Benford’s Law as a consequence of entropy. Related discussions included mystifying concepts such as Shannon entropy, emitters and receivers, binary files, and other inexplicable ideas in Information Theory — a field which Kafri conjectures to correspond to thermodynamics. A long chapter in the 5th section of the book is dedicated solely to his model.

I warmly thank Lissette Picado for straightening out and polishing some of my long, winding, and complex sentences, and for saving others from outright ruin

and grammatical distortion — bringing to the rescue superb English language ability and commendable professionalism.

I warmly thank Adrian Saville of South Africa for carefully reviewing the chapters pertaining to his data compliance test known in this book as ‘Saville Regression Algorithm’, as well as for reviewing other parts of the book before publication. I am indebted to him for coming up with this innovative and elegant measure of digital configuration, a measure which turned out to be exceedingly useful in concisely measuring Digital Development Pattern.

I thank professor of accountancy and auditing Charles Carslaw of New Zealand (currently at the University of Nevada in Reno, USA) for reviewing the relevant sections on fraud detection, forensic digital analysis, and other topics, and sending me back many useful comments before publication. As evident from the date on Carslaw’s groundbreaking article in 1988 utilizing the 2nd order digital distribution to [successfully] detect fraudulent accounting and financial reporting by New Zealand firms, Carslaw is the first person credited with publishing an application of Benford’s Law in the context of fraud detection and forensic digital analysis regarding financial data. Carslaw’s original insight was to compare digital configuration of financial data with the Benford configuration, and to suggest the possibility of [industry-wide] fraud if discrepancy is deemed significant. Quoting from <http://www.audimation.com/pdfs/guide-to-benford-s-law.pdf>: “Benford himself saw no practical benefit from his work. It was not until 1988 that Benford’s Law was cited in a survey by Charles Carslaw focusing on the second digit in a list of company earnings. Carslaw used the frequencies calculated by Benford as a benchmark for the result of his analyses.” A similar revelation occurred to Hal Varian — chief economist at Google Inc. — who in 1972 gave the first ever application of the law in algorithm error detection, and came quite close to suggesting such an application of the law for the detection of intentional human manipulation of data. In communicating with me, Varian wrote that even though he only considered Benford’s Law as a check for error, and did not formally mention fraud explicitly, nonetheless he had no particular view about whether the error was intentional or not (i.e., whether it was fraud). Other digital analysts and accountants who contributed to the field of fraud detection in the context of Benford’s Law are mentioned and acknowledged in the second section. I thank Varian, Carslaw, and all those who contributed to fraud detection for the knowledge and insight they provided me, and for their direct or indirect contribution to this book.

Finally, I would like to pay tribute and express my debt to my beloved late father Shaull Kossovsky — a great scholar and the author of several acclaimed books — for stressing the value of knowledge and learning for its own sake, and for the inspiration and influence bequeathed me.

I wish to acknowledge my appreciation and debt to the Republic of Costa Rica [where I am currently staying] for its peaceful existence, having abolished its military following a rigged election in 1948 with the malevolent involvement of its armed forces. It is almost a certainty that the digital structure of the vote tally in that fraudulent election was quite contrary to the Benford configuration, and thus could have been easily detected forensically. Living in a country without a single menacing soldier, tank, or artillery piece, is quite unique and refreshing, and exceedingly rare in this troubled world. In this aspect Costa Rica is a light unto the nations, an essential example of how we all should live. In the age of horrific weapons, numerous [unnecessary and invented] international conflicts, renewed talk about possible nuclear wars, and worries about the annihilation of our entire civilization and species, Costa Rica leads by example, and points the way forward. I also wish to express gratitude to Costa Rica for hosting me while I was creating and writing the latter half of this book, including the entire 5th and 7th sections, and parts of the 4th and the 6th. This book will always be associated in my mind with that small and peaceful nation caught between two large and intimidating oceans.

# CONTENTS

---

---

<i>Benford's Law</i>	v
<i>Foreword</i>	vii
<i>Introduction</i>	ix
<i>Acknowledgment</i>	xiii
<b>Section 1: Benford's Law</b>	<b>1</b>
1. Digits versus Numbers	3
2. To Find Fraud, Simply Examine Its Digits!	5
3. First Leading Digits	8
4. Empirical Evidence from Real-Life Data on Digit Distribution	9
5. Physical Clues of the Digital Pattern	15
6. Historical Background of the Two Discoverers	18
7. Benford's Law	21
8. The Prevalence of Benford's Law	29
9. Physical Law versus Numerical Law	31
10. Nature's Way of Counting Single-Issue Phenomena	33
11. Case Study I: Time Between Earthquakes	38
12. Data on Population Counts of Cities, Towns, Regions, and Districts	41
13. Case Study II: U.S. Census Data on Population Centers	42
14. Data sets on USA Population by State and by County	46
15. Four Distinct Numerical Processes Leading to Benford	48
16. Random Linear Combinations and Accounting Revenue Data	49
17. Aggregation of Data Sets as a Prominent Cause of Benford's Law	53
18. Random Pick from a Variety of Data Sources is Logarithmic	55
19. Integral Powers of Ten	57
20. The Logarithmic as Repeated Multiplications	58
21. Case Study III: Exponential 0.5% Growth Series for 3,233 Periods	67
22. Case Study IV: 140 Cumulative Dice Multiplications	70
23. The Universality of Benford's Law — True in any Scale System	72

24. A Hidden Digital Signature within Benford's Digital Signature	74
<b>Section 2: Forensic Digital Analysis &amp; Fraud Detection</b>	<b>77</b>
25. Historical Background of the First Applications of Benford's Law	79
26. Methods in Financial and Accounting Fraud Detection	81
27. The Part and Type of Data Applicable to Forensic Testing	88
28. Case Study V: U.S. Market Capitalization on January 1, 2013	96
29. Case Study VI: Microsoft Corporation Financial Statement	98
30. Case Study VII: Total Return of Athena Guaranteed Futures Fund	100
31. Establishing Direct Connection Between Digit Anomaly & Fraud	102
32. Post-Test Conclusions	106
33. Detecting Fraud via Digital Development Pattern	108
34. The Dilemma of FTD versus LTD for Digit-Anemic Numbers	110
<b>Section 3: Data Compliance Tests</b>	<b>113</b>
35. Testing Data for Conformity to Benford's Law	115
36. The Z Test	120
37. The chi-Square Test	123
38. SSD as a Measure of Distance from the Logarithmic	128
39. Saville Regression Measure	134
40. Value Repetition Test	138
41. The Confusion and Mistaken Applications of Summation Test	141
42. Summation Test in the Context of Fraud Detection	147
43. Methods in Digital Development Pattern Detection	149
44. Case Study VIII: Price List of a Large Manufacturer	164
45. Case Study IX: USA County Area Data	172
46. Random Linear Combinations and Revenue Data Revisited	177
47. Case Study X: Forensic Analysis of Revenue Data for Small Shop	192
<b>Section 4: Conceptual and Mathematical Foundations</b>	<b>195</b>
48. Hybrid Data Sets Blending Several Data Types	197
49. Second-Generation Distributions	198
50. A Leading Digits Parable	200
51. Simple Averaging Scheme as a Model for Typical Data	207
52. More Complex Averaging Schemes	212
53. Digital Proportions within the Number System Itself	216

54. Chains of Distributions	219
55. Hill's Super Distribution	227
56. The Scale Invariance Principle	230
57. Philosophical and Conceptual Observations	233
58. Some General Results	237
59. Density Curves and Their Leading Digits Distributions	242
60. The Case of $k/x$ Distribution	246
61. Uniform Mantissa, Varied Significant, and the General Law	253
62. Uniqueness of $k/x$ Distribution	261
63. Related Log Conjecture	266
64. Testing Related Log Conjecture via Simulations	271
65. The Lognormal Conjecture of Hill's Super Distribution	275
66. Non-Symmetric Related Log Curves	279
67. Wide Range on the Log-Axis and Logarithmic Behavior	281
68. The Remarkable Malleability of Related Log Conjecture	282
69. Hill's Super Distribution and Related Log Conjecture	293
70. Scale Invariance Principle and Related Log Conjecture	295
71. The Near Indestructibility of Higher Order Distributions	297
72. Falling Density Curve with a Tail to the Right	301
73. Falling Density Curve with a Particular Steepness	305
74. Fall in Density is Well-Coordinated Between IPOT Values	307
75. Synthesis Between the Deterministic and the Random	313
76. Dichotomy Between the Deterministic and the Random	317
77. Fitting the Random into the Deterministic	327
78. The Random Flavor of Population Data	332
79. The Lognormal Distribution and Benford's Law	335
80. Scrutinizing Digits within Lognormal, Exponential, and $k/x$	339
81. Leading Digits Inflection Point	345
82. Digital Development Pattern Found in all Real-Life Random Data	349
83. Digital Development Pattern Seen Only Under IPOT Partition	356
84. Development Pattern More Prevalent than Benford's Law Itself	360
85. Sum-Invariant Characterization of the Law (Summation Test)	363
<b>Section 5: Benford's Law in the Physical Sciences</b>	<b>373</b>
86. Mother Nature Builds and Destroys with Digits in Mind	375
87. Quantum Mechanics, Thermodynamics, and Benford's Law	377

88.	Chemistry, Random Linear Combinations, and Benford's Law	380
89.	Benford's Law and the Set of all Physical Constants	387
90.	MCLT as an Explanation for Single-Issue Physical Phenomenon	389
91.	Chains as an Explanation for Single-Issue Physical Phenomenon	395
92.	Breaking a Rock Repeatedly into Small Pieces is Logarithmic	398
93.	Random Throw of Balls into Boxes Approximating the Logarithmic	402
94.	Logarithmic Model for Planet and Star Formations	415
95.	Hybrid Causes Leading to Logarithmic Convergence	419
96.	Mild Deviations Seen in Small Samples of Logarithmic Data Sets	421
97.	The Remarkable Versatility of Benford's Law	423
<b>Section 6: Topics in Benford's Law</b>		<b>425</b>
98.	Singularities in Exponential Growth Series	427
99.	Super Exponential Growth Series	439
100.	Higher-Order Leading Digits	442
101.	Digit Distributions Assuming Other Bases	450
102.	Chains of Distributions Revisited	452
103.	Chainable Distributions and Parameters	462
104.	Frank Benford's Averaging Scheme as a Distribution Chain	471
105.	Effects of Parametrical Transformations on Leading Digits	474
106.	Digits of the Wald, Weibull, chi-square, and Gamma Distributions	479
107.	Digital Patterns of the Exponential Distribution	480
108.	Saville Regression Measure Revisited	484
109.	The Scale Invariance Principle and AGD Interpretation	497
110.	Case Study XI: Large Sample from a Variety of Data Sources	501
111.	Direct Expression of first Digit for any Number — Computer Use	506
112.	Artificially Creating Nearly Perfect Logarithmic Data	507
<b>Section 7: The Law of Relative Quantities</b>		<b>509</b>
113.	The Relating Concepts of Digits, Numbers, and Quantities	511
114.	Benford's Law in its Purest Form	513
115.	Number System Invariance Principle	519
116.	Cartesian Coordinate System is Number-System-Invariant	522
117.	Physics is Number-System-Invariant	524
118.	Multiplicative CLT is Number-System-Invariant	527
119.	Greek Parable and Chains are Number-System-Invariant	529

120.	Physical Reality versus Digital Perception	530
121.	Patterns in Physical Data Transcend Number Systems and Digits	531
122.	Common Thread Going Through Multiple Physical Data Sets	532
123.	Casting a Repetitive Bin System to Measure Fall in Histogram	535
124.	Non-Expanding Bin System Measuring Fall in $k/x$ Distribution	542
125.	Once-Expanding Bin System Measuring Fall in $k/x$ Distribution	547
126.	Once-Expanding Bins for $k/x$ Reduces to Benford when $F = D + 1$	549
127.	Twice-Expanding Bin System Measuring Fall in $k/x$ Distribution	550
128.	Twice-Expanding Bins for $k/x$ Reduces to Benford when $F = D + 1$	552
129.	Infinitely Expanding Bin System Measuring Fall in $k/x$	554
130.	Confirmation Matching $k/x$ Fall with Empirical Bins on Real Data	556
131.	Closed Form Expression for the Limit of the Infinite Sequence	558
132.	Closed Form Expression for the Limit in the Flat Case $F = 1$	563
133.	9-Bin Systems with $F = 10$ on Real Data All Yield $\text{LOG}_{\text{TEN}}(1+1/d)$	567
134.	Bin Systems Need to Start Near Origin with Small Initial Width	569
135.	Actual or Degree of Compliance May Be Bin- and Base-Variant	575
136.	Correspondence in Data Classification Between Bin Systems and BL	582
137.	$F = D + 1$ Bin Systems on Real Data Yield $\text{LOG}_{\text{BASE}}(1+1/d)$	590
138.	The Remarkable Malleability and Universality of Bin Schemes	591
139.	Higher-Order Digits Interpreted as Particular Bin Schemes	602
140.	Bin Development Pattern	608
141.	The General Scale Invariance Principle	614
142.	Paradoxes Explained	619
143.	Digits Serving as Quantities in Benford's Law	622
144.	Frank Benford's Prophetic Words	624
145.	Future Direction	625
146.	The Universal Law of Relative Quantities	626
147.	Dialogue Concerning the Two Chief Statistical Systems	628
	<i>References</i>	637
	<i>Glossary of Frequently Used Abbreviations</i>	643
	<i>Index</i>	645

**This page intentionally left blank**

# **Section 1**

## **BENFORD'S LAW**

**This page intentionally left blank**

## DIGITS VERSUS NUMBERS

---

---

The typical statistician, during a typical day at the office, spends most of the time intensely staring at data charts and scatter plots, seeking real or imaginary patterns where perhaps none exist, summarizing data, calculating averages and standard deviations, regressing and correlating seemingly unrelated variables, analyzing subtle variances between related data sets to determine whether they are significantly or randomly different from each other, dissecting and bisecting those pesky numbers sent by clients, government agencies, companies, and research institutes.

Interestingly, the statistician is recently taking on the role of a philosopher of sorts, and instead of examining the numbers themselves as is the standard practice, he or she is investigating the digital language utilized in writing those numbers. What letters are to words, digits are to numbers. Why should a poetry lover seek any patterns or beauty by looking into the letters in Shakespeare's prose instead of the elegantly combined words? Yet, the relative proportions of our ten digits 0 to 9 occurring within our typical everyday numbers are now being routinely recorded and investigated by statisticians and data analysts, and even theorized as to how exactly they should be spread within any given data set by applying mathematical and statistical reasoning. Moreover, the study of digit proportions is further subdivided by classifying them into different categories according to position. For example, the specific proportions of the leftmost digit, namely the first digit of numbers, is looked into and examined separately. Another separate analysis is performed on the second-leftmost digit, which indeed shows quite different digital proportions than those of the first digit. But aren't all digits supposed to be occurring randomly and thus equally distributed? Why should the digit 4 for example have a higher or lower chance of occurring within numbers than say the digit 5? One wonders whether the occurrences of digits themselves within numbers are just '*too random*' for the statistician to even consider and analyze. Is there indeed a particular statistical law supposedly governing digital proportions? In addition, it seems doubtful that there would be any use or consequence in

looking into this digital language proportion in the first place. Are there any applications that can exploit the examination of these digital proportions?

The answers to the latter two questions are all decisively positive, as evident by the newly-created role assigned to the statistician recently as a private detective utilizing known digital patterns in data to detect fraud by knowing that fake data probably lacks those particular digital patterns. Previously, the task of the statistician was merely to analyze data, but never to decide on the authenticity of the provided data. Data was traditionally always taken as a given without any ability to authenticate. For how could the unsuspecting, honest and naive statistician know that people were sending him or her fake data that was merely invented? One incentive to fake data and reduce reported revenues and income would naturally be to lower tax payments. Another incentive is the temptation to inflate revenues and profits in order to impress investors and present the company in a better light as being financially sound. Therefore there is a strong need on the part of tax authorities, governmental financial regulatory and supervisory agencies worldwide, as well as auditing and accounting companies and others, to obtain professional statistical advice as to how to detect fake data. By wearing that philosopher's hat and examining the digital language used in writing the numbers in provided data sets, the statistician is then able to wear his or her other hat, namely the detective's hat, and forensically analyze data for any possible fraud.

## TO FIND FRAUD, SIMPLY EXAMINE ITS DIGITS!

---

---

As our civilization progresses, we are able to do things previously thought impossible. Our collective mathematical and technological abilities have reached fantastic heights. We literally perform magic with our computers and other gadgets. But can we perform the simple task of telling when a friend or a spouse lies? Perhaps not, but the truly sophisticated statistician, aware of the latest developments in the field, can nowadays detect straight-faced fraudsters when presented with their fake data. Underpinning this ability is the fact that to concoct authentic-looking data one must know something about the particular properties of their digital language, while most fraudsters haven't got a clue about the topic, and mistakenly believe that digital equality rules the universe of numbers. Yet in fact, low digits such as 1, 2, and 3 actually occur with very high frequencies within the first-place position of typical everyday data, while high digits such as 7, 8, and 9 have very little overall proportion of occurrence. So much so that the proportion of everyday typical numbers starting with digit 1 is about seven times that of numbers starting with digit 9! About 30% of typical everyday numbers in use start with digit 1, while only about 4% start with digit 9.

In order to illustrate the ability of utilizing this peculiar digital phenomenon in fraud detection, we shall digitally analyze hypothetical accounting data from five different companies where amounts represent revenues. The table in Fig. 1.1 shows 25 dollar amounts from each company. Nothing seems unusual or suspicious if we merely focus on the numbers themselves. Yet, if we forensically investigate the digital language used in writing those numbers, namely the digits at the very beginning of each number (the leftmost ones), we can immediately reveal an abnormality with one particular data set. Figure 1.2 shows the proportions of the first digits for all five companies.

Clearly, MF Capital comes under strong suspicion in the eyes of the expert statistician, since typical accounting data rarely comes with anything near digital equality for the first position. First-digit proportions of the other four companies show an overall pattern of gradual decrease, consistent with the expected pattern

Alcoa	Amgen	Motorola	MF Capital	Xerox
624.00	120.30	182.03	420.40	13.00
149.00	74.40	158.11	74.00	62.69
104.74	107.71	37.62	3.54	69.73
171.00	17.43	363.60	645.00	221.05
102.26	9.99	361.99	211.40	57.01
179.98	373.68	150.00	8.40	98.00
9.77	209.00	209.87	143.24	11.25
373.87	14.09	250.61	23.87	56.94
23.48	54.00	3.62	5.64	225.00
12.98	219.00	575.44	503.24	596.76
480.61	3.62	79.45	978.20	4.99
4.37	636.20	245.53	10.87	44.96
116.25	1332.81	404.84	43.80	493.05
149.00	174.38	86.77	3.84	120.84
274.92	32.40	114.35	8.97	97.04
89.00	225.00	119.89	935.70	11.42
20.70	78.00	19.10	7.25	100.00
224.93	479.00	6.62	54.30	48.00
50.00	3.62	364.00	29.35	11.98
662.75	529.75	25.10	74.60	50.00
109.90	63.60	85.35	631.70	20.00
842.98	30.26	22.00	6.54	24.09
403.00	1706.63	16.50	4.94	160.00
383.50	124.00	154.46	36.70	270.00
1494.41	1328.11	55.12	2.98	6.35

Figure 1.1 Hypothetical Accounting Data for Five Companies

in almost all types of accounting data. The set of the first digits for MF Capital revenue data (commas omitted) is  $\{4736281255914389752766432\}$ , which is distinctly different compared to say Alcoa's  $\{6111119321441128225618431\}$ . Digits at the second and third positions are much more equal in proportions for all five companies and do not show any particular pattern; they also do not single out MF Capital in any way. Had the focus of the statistician been misplaced on those digits, there wouldn't be any clue about MF Capital's possible fraudulent activities.

<b>Digit</b>	<b>Alcoa</b>	<b>Amgen</b>	<b>Motorola</b>	<b>MF Capital</b>	<b>Xerox</b>
<b>1</b>	<b>40%</b>	<b>36%</b>	<b>32%</b>	<b>8%</b>	<b>28%</b>
<b>2</b>	<b>16%</b>	<b>12%</b>	<b>20%</b>	<b>16%</b>	<b>20%</b>
<b>3</b>	<b>8%</b>	<b>20%</b>	<b>20%</b>	<b>12%</b>	<b>0%</b>
<b>4</b>	<b>12%</b>	<b>4%</b>	<b>4%</b>	<b>12%</b>	<b>16%</b>
<b>5</b>	<b>4%</b>	<b>8%</b>	<b>8%</b>	<b>12%</b>	<b>16%</b>
<b>6</b>	<b>8%</b>	<b>8%</b>	<b>4%</b>	<b>12%</b>	<b>12%</b>
<b>7</b>	<b>0%</b>	<b>8%</b>	<b>4%</b>	<b>12%</b>	<b>0%</b>
<b>8</b>	<b>8%</b>	<b>0%</b>	<b>8%</b>	<b>8%</b>	<b>0%</b>
<b>9</b>	<b>4%</b>	<b>4%</b>	<b>0%</b>	<b>8%</b>	<b>8%</b>

**Figure 1.2** 1st Digits Proportions of the Data of Five Companies

## FIRST LEADING DIGITS

---

---

First Leading Digit (LD) or First Significant Digit is the first (non-zero) digit of a given number appearing on the leftmost side. For 567.34 the leading digit is 5. For 0.0367 the leading digit is 3, as we discard the zeros. For the lone integer 6 the leading digit is 6. For negative numbers we simply discard the sign, hence for -62.97 the leading digit is 6. Another way of defining the first digit of any number is by writing it in scientific notation as  $A \cdot 10^N$  with  $N$  being an integer and  $A$  being a real number such that  $1 \leq |A| < 10$ . For such representation of numbers, the integral part of  $A$  (excluding the fractional part), and with the positive or negative sign ignored, is what we consider the first leading digit. For example, the number 311.75 is scientifically written as  $3.1175 \cdot 10^2$  and digit 3 leads the number. Naturally, when digit  $d$  appears first in a number composed of several digits, we call  $d$  the **'leader'**, as it leads all the other digits trailing behind it to the right.

613             $\longrightarrow$  digit 6

0.0002867    $\longrightarrow$  digit 2

7               $\longrightarrow$  digit 7

-7              $\longrightarrow$  digit 7

1,653,832    $\longrightarrow$  digit 1

-0.456398    $\longrightarrow$  digit 4

## EMPIRICAL EVIDENCE FROM REAL-LIFE DATA ON DIGIT DISTRIBUTION

---

---

Perhaps it is tempting to intuit that for numbers in typical real-life data sets, all nine digits  $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$  should be equally likely to occur and thus uniformly distributed. Let us examine three typical data sets from a variety of real-life situations where digital results run counter to that misguided intuition and where, surprisingly, low digits such as 1, 2, and 3 are strongly favored over high digits such as 7, 8, and 9. The three data sets to be digitally examined are: (I) stock market prices and volume of stock traded, (II) the 10 by 10 multiplication table, and (III) house number in typical address data.

Examination of first digits of closing prices and daily volume of stocks traded on the New York Stock Exchange on December 23, 2011 reveals a definite pattern in which digital proportions are almost monotonically and consistently decreasing. The first 31 companies on top of the alphabetically-sorted list were arbitrarily chosen. Figure 1.3 shows the extracted data.

Low digits lead much more often than high digits, for both stock prices and volume. Figure 1.4 shows the exact LD distributions for this limited set of 31 companies. It should be noted that almost all other such subsets down the long list on the NYSE website yield quite similar results, that there was nothing unusual about the trading day of the 23rd of December 2011, and that very similar digital results are gotten on other trading days.

Let us examine LD of the 10 by 10 multiplication table that we all were forced to memorize at school against our will, as shown in Fig. 1.5(A).

Surprisingly, out of 100 numbers, 21 start with the *lowest* digit 1 (shown in large and bold font), and only five start with the *highest* digit 9 (shown within circles), namely a ratio of 4:1 roughly. This result is surprising yet approximately compatible with the digital results seen in the example with stock prices and volume data. In this digital analysis the numbers 1, 10, and 100 are grouped together under the same category since all of them are being led by digit 1. Digital proportions here are  $\{21\%, 17\%, 13\%, 14\%, 8\%, 9\%, 6\%, 7\%, 5\%\}$ .

Stock Symbol	Closing Price	NYSE Volume
A	\$ 30.74	1,124,700
AA	\$ 38.32	5,950,900
AAI	\$ 7.03	533,700
AAP	\$ 34.09	430,100
AAR	\$ 22.14	8,600
AAV	\$ 11.01	263,800
AB	\$ 60.86	335,400
ABA	\$ 25.75	4,000
ABB	\$ 25.12	2,627,700
ABC	\$ 41.48	478,600
ABD	\$ 14.03	264,200
ABG	\$ 14.24	164,500
ABH	\$ 9.68	992,700
ABI	\$ 34.42	791,000
ABK	\$ 9.94	1,688,700
ABM	\$ 19.88	140,100
ABN	\$ 57.62	29,500
ABN PRE	\$ 21.49	28,000
ABN PRF	\$ 23.58	5,800
ABN PRG	\$ 22.15	46,100
ABR	\$ 15.92	254,700
ABT	\$ 53.23	2,336,000
ABV	\$ 85.17	406,200
ABV C	\$ 77.19	5,400
ABW PRA	\$ 25.02	1,900
ABX	\$ 53.55	2,574,500
ACC	\$ 26.52	147,300
ACE	\$ 55.09	1,216,700
ACE PRC	\$ 24.92	11,300
ACF	\$ 14.50	597,600
ACG	\$ 8.39	193,300

**Figure 1.3** Price and Volume of Stocks Traded on the NYSE

Interestingly, if the digital aspect of the multiplication table is ignored and the focus shifts to pure quantities, the result is still quite similar! The entire range of (1, 100) is partitioned into 10 equitable sections (1, 10), (11, 20), (21, 30), (31, 40), (41, 50), (51, 60), (61, 70), (71, 80), (81, 90), (91, 100), and a count is made of the numbers falling within each section — namely grouping them according to quantities. Figure 1.5 (B) demonstrates such quantitative partitioning of the entire territory in detail. Figure 1.5 (C) gives the counts of numbers falling within each section, where quantitative proportions are {27%, 19%, 15%, 11%, 9%, 6%, 5%, 4%, 3%, 1%}. Clearly there are numerous

1st Digit	Price	Volume
1	19.4%	29.0%
2	29.0%	25.8%
3	12.9%	3.2%
4	3.2%	16.1%
5	12.9%	16.1%
6	3.2%	0.0%
7	6.5%	3.2%
8	6.5%	3.2%
9	6.5%	3.2%

Figure 1.4 1st Digits of Stock Price & Volume

*	1	2	3	4	5	6	7	8	9	10
1	<b>1</b>	2	3	4	5	6	7	8	9	<b>10</b>
2	2	4	6	8	<b>10</b>	<b>12</b>	<b>14</b>	<b>16</b>	<b>18</b>	20
3	3	6	9	<b>12</b>	<b>15</b>	<b>18</b>	21	24	27	30
4	4	8	<b>12</b>	<b>16</b>	20	24	28	32	36	40
5	5	<b>10</b>	<b>15</b>	20	25	30	35	40	45	50
6	6	<b>12</b>	<b>18</b>	24	30	36	42	48	54	60
7	7	<b>14</b>	21	28	35	42	49	56	63	70
8	8	<b>16</b>	24	32	40	48	56	64	72	80
9	9	<b>18</b>	27	36	45	54	63	72	81	90
10	<b>10</b>	20	30	40	50	60	70	80	90	<b>100</b>

Figure 1.5 (A) 1st Digits Comparison within Multiplication Table

small quantities within the multiplication table but very few big quantities, and therefore it is skewed quantitatively as well as digitally. It is highly unlikely that the digital phenomenon drives the quantitative phenomenon. Most probably the quantitative drives the digital. It would be very difficult to believe that by some magical and rare coincidence there are actually two distinct and independent phenomena out there, (I) digital — favoring low digits, and (II) quantitative — favoring low quantities.

*	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10
2	2	4	6	8	10	12	14	16	18	20
3	3	6	9	12	15	18	21	24	27	30
4	4	8	12	16	20	24	28	32	36	40
5	5	10	15	20	25	30	35	40	45	50
6	6	12	18	24	30	36	42	48	54	60
7	7	14	21	28	35	42	49	56	63	70
8	8	16	24	32	40	48	56	64	72	80
9	9	18	27	36	45	54	63	72	81	90
10	10	20	30	40	50	60	70	80	90	100

Figure 1.5 (B) Quantitative Territorial Partitioning of the Multiplication Table

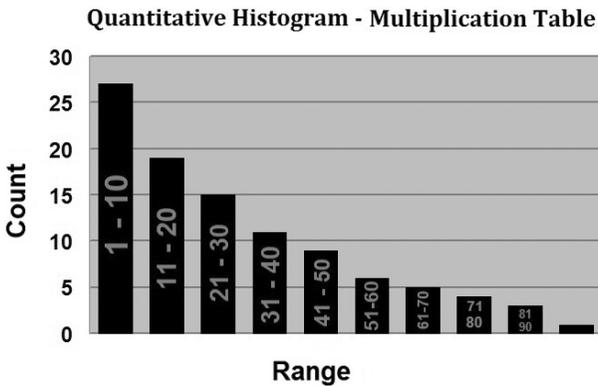


Figure 1.5 (C) Histogram of Relative Quantities — Multiplication Table

Could the multiplication table serve as a good representative of real-life data or is it merely a calculation tool and its results are irrelevant? To see how the table could reflect real data sets or processes, a game is imagined where two virtual dice having 10 sides each are thrown simultaneously. The casino attempting to attract

the mathematically-inclined and more sophisticated gamblers provides this game, where large sums of money are bet that the value of the product of the two dice is over 50 (*which on the face of it — is reasonable and fair to all — being that 50 is the midpoint of the entire range of possibilities 1–100*). The rule for this highly popular and addictive game is that the shrewd casino owner wins on small quantities 1–50, while naïve gamblers win on big quantities 51–100, enabling the casino to earn a steady income every night. This scenario demonstrates how our deterministic 10 by 10 table could metamorphose into the random. As it happened, this process, namely ‘**multiplying the randoms**’, is actually highly relevant to many physical processes and real-life data sets.

Finally we examine digital proportions of a small sample of 21 street addresses randomly chosen from the Yellow Pages in South Dakota, USA, as shown in Fig. 1.6. The focus here is on the **house number**, not on the street number or the zip code. This small segment of data from the much larger population data set was taken strictly in a random fashion without attempting to be selective in order to achieve any particular digital configuration. Similar results are obtained over

4908	S Glenview Rd, Sioux Falls, SD, 57108
2516	Clarkway Dr, Sioux Falls, SD, 57105
403	West 11th St, Canton, SD, 57013
27348	461st Ave, Chancellor, SD, 57015
316	S Potter St, Gettysburg, SD, 57442
18769	Quin Rd, Nisland, SD, 57762
638	S. Main Ave, Apt. 10, Sioux Falls, SD, 57104
273	Minnesota St, Rapid City, SD, 57701
3801	S Judy Ave, Sioux Falls, SD, 57103
504	Ironwood Dr, Hartford, SD, 57033
11853	391 Ave, Columbia, SD, 57433
2902	Tomahawk Dr, Rapid City, SD, 57702
13687	387th Ave, Aberdeen, SD, 57401
603	2nd Ave West, Flandreau, SD, 57028
38485	129th St, Aberdeen, SD, 57401
13497	465th Ave, Wilmot, SD, 57279
43404	188th St, Willow Lake, SD, 57278
1010	Valley View Court, Huron, SD, 57350
312	Alta Vista Dr, Rapid City, SD, 57701
111	Lee Hill Rd, Pierre, SD, 57501
648	13th St SW, Huron, SD, 57350

Figure 1.6 A Sample of Address Data from South Dakota, USA

and over again from almost any other segment of the Yellow Pages. Here again, low digits lead strongly, taking by far the lion's share of overall proportion. In fact, the set of first digits is {424231623512163141316}. Consequently digit proportions here are {28.6%, 19%, 19%, 14.3%, 4.8%, 14.3%, 0%, 0%, 0%}, excluding digit 7, 8, and 9 altogether, while allocating digit 1 nearly a third of total leadership!

In conclusion: approximately the same overall digital pattern was found in all three real-life examples above, strongly suggesting that first digits are not at all equally distributed.

## PHYSICAL CLUES OF THE DIGITAL PATTERN

---

---

Simon Newcomb in 1881, and then Frank Benford independently in 1938, discovered the law governing digital distributions in typical real-life data. One wonders what might have motivated them to look into the subject matter in the first place. Interestingly, both were inspired by the same phenomenon. Both start their articles by mentioning an odd observation about some common appearances of old logarithm books. Almost nobody noticed in the old days before the advent of the hand-held calculator and the computer that books of tables of logarithms, square roots, trigonometric values, and such, were almost always much more worn out at the beginning pertaining to numbers that start with digit 1 or 2 than at the end pertaining to numbers that start with digit 8 or 9. That wear and tear of the physical pages seemed to negatively correlate with the first digit, since it was progressively less severe throughout the book for higher digits. Naturally Newcomb and Benford took it as evidence that people, engineers, and scientists were on average more in need of using the first pages than the last pages, reflecting the spread of first digits in the real world. This consistent variation in the use of the pages according to first-digit value in almost all such books constituted the first hint of the phenomena and led Newcomb and Benford to discover and then to investigate the pattern. Harry Benford distinctly recalls his father saying that it was exactly these variations in usage of old logarithm books that caught his attention.

Newcomb's first sentence refers to this evidence. *"That the ten digits do not occur with equal frequency must be evident to anyone making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones."* Benford begins his article with elaborated reasoning about what logarithm books may ultimately hint at. *"It has been observed that the pages of a much used table of common logarithms show evidences of a selective use of the natural numbers. The pages containing the logarithms of the low numbers 1 and 2 are apt to be more strained and frayed by use than those of the higher numbers 8 and 9. Of course, no one could be expected to*

A TABLE  
CONTAINING THE  
LOGARITHMS OF NUMBERS  
FROM 1 TO 10,000

---

No.	Log.								
1	0.000000	21	1.322219	41	1.612784	61	1.785330	81	1.908485
2	0.301030	22	1.342423	42	1.623249	62	1.792392	82	1.913814
3	0.477121	23	1.361728	43	1.633468	63	1.799341	83	1.919078
4	0.602060	24	1.380211	44	1.643453	64	1.806180	84	1.924279
5	0.698970	25	1.397940	45	1.653213	65	1.812913	85	1.929419
6	0.778151	26	1.414973	46	1.662758	66	1.819544	86	1.934498
7	0.845098	27	1.431364	47	1.672098	67	1.826075	87	1.939519
8	0.903090	28	1.447158	48	1.681241	68	1.832509	88	1.944483
9	0.954243	29	1.462398	49	1.690196	69	1.838849	89	1.949390
10	1.000000	30	1.477121	50	1.698970	70	1.845098	90	1.954243
11	1.041393	31	1.491362	51	1.707570	71	1.851258	91	1.959041
12	1.079181	32	1.505150	52	1.716003	72	1.857332	92	1.963788
13	1.113943	33	1.518514	53	1.724276	73	1.863323	93	1.968483
14	1.146128	34	1.531479	54	1.732394	74	1.869232	94	1.973128
15	1.176091	35	1.544068	55	1.740363	75	1.875061	95	1.977724

Figure 1.7 Logarithm Table from the Pre-Computer Era

*be greatly interested in the condition of a table of logarithms, but the matter may be considered more worthy of study when we recall that the table is used in the building up of our scientific, engineering, and general factual literature”.*

Figures 1.7 and 1.8 are two pictures of old logarithmic books. The logarithm table in Fig. 1.7 contains a tantalizing clue of the exact logarithmic proportion 0.301030 for digit 1 as predicted by Benford's Law — namely 30.1% chance of finding digit 1 leading a typical number in real-life data sets, as shall be discussed later.

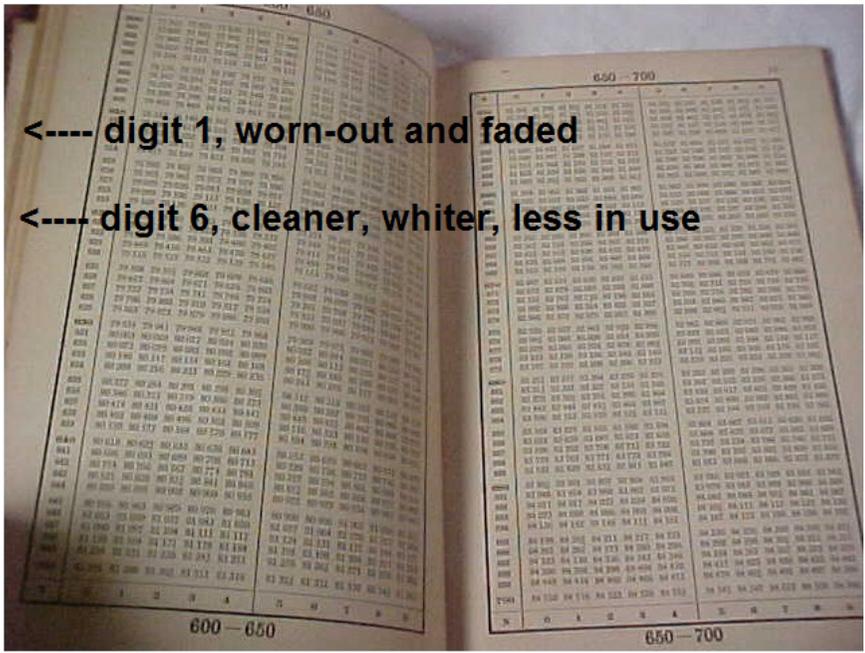


Figure 1.8 Differentiation in Wear and Tear Hinting at the Digital Phenomena

## HISTORICAL BACKGROUND OF THE TWO DISCOVERERS

---

---

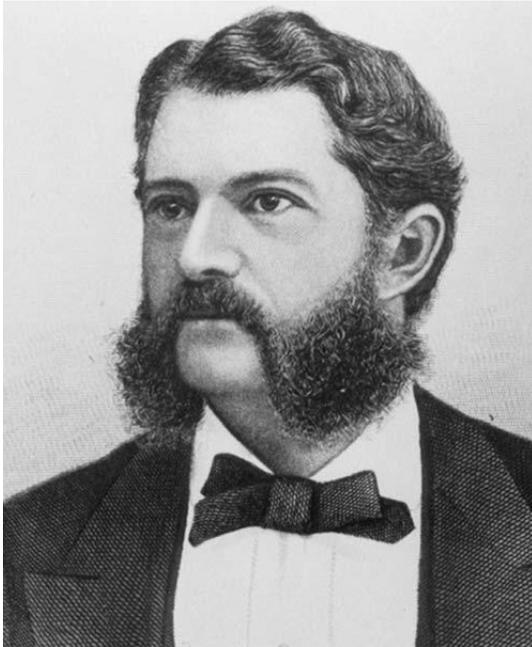
**Simon Newcomb** (1835–1909) was a Canadian–American astronomer and mathematician known for his talent in dealing with immense amounts of numerical data. Though he was self-taught and had almost no conventional schooling as a boy, he made important contributions to astronomy and physics. Working at the United States Naval Astronomical Observatory, and later becoming the director of the US Navy’s Nautical Almanac, Newcomb set out to work on the measurement of the position of the planets and the moon as an aid to navigation, becoming increasingly interested in theories of planetary motion. In 1878, Newcomb had started planning for a new and precise measurement of the speed of light that was needed to account for exact values of many astronomical constants. He had a long collaboration and friendship with Albert Michelson, who was also working on measurement projects regarding the speed of light, and who was failing to detect any supposed absolute motion of Earth through that mythical medium termed the ‘aether’, hinting at Relativity. By 1882 Newcomb had recalculated Mercury’s Perihelion Motion, obtaining by far more accurate values for the small ‘flaw’ in Mercury’s orbit around the Sun as predicted by Newton’s laws. Einstein’s predictions of Mercury’s orbit almost exactly matched Newcomb’s careful observations, and this served as one of the strongest confirmation of General Relativity at the time of its inception in 1915.

While Newcomb’s various contributions paved some of the harsh terrain to Relativity and were widely acclaimed and acknowledged, almost no one noticed his two-page article *Note on the Frequency of Use of the Different Digits in Natural Numbers* published in 1881 where he correctly described and tabulated the digital distributions known today as Benford’s Law. Without explicitly writing an analytical expression for the proportions of the digits, Newcomb nonetheless mathematically stated the law indirectly in complete generality. About a page of his article was devoted to a failed attempt at explaining the phenomena by claiming that typical numbers in nature occur as ratios of two distinct quantities. His article

was subsequently forgotten and ignored for almost six decades until Benford aroused interest in the phenomenon with his acclaimed article in 1938.

**Frank Benford** (1883–1948) was an American physicist and electrical engineer. He worked in General Electric, published 109 papers in the fields of optics and mathematics, and was granted 20 patents on optical devices. Benford, who had no knowledge of Newcomb's earlier work, independently re-discovered the phenomenon, and in 1938 wrote a lengthy article about it titled *The Law of Anomalous Numbers*. As opposed to Newcomb who for the most part just stated the fact in a broad mathematical statement without offering forensic evidence, Benford explicitly stated the algebraic expression for the probability of the first digits, and set out to empirically examine 20 different large collections of data types, systematically recording their digital results and compatibility with the law.

Benford examined a variety of different data sets, such as area of rivers, population census data, specific heat of chemical substances, atomic weights, house address data, some mathematical sequences, including interestingly a totally random pick of numbers from newspapers. Agreement with the law was quite



**Figure 1.9** Simon Newcomb in 1871, (Image credit: U.S. Naval Observatory)



**Figure 1.10** Frank Benford

strong, especially for the resultant average of all his 20 data sets. Benford's failed attempt at explaining the phenomenon mathematically involved a particular averaging model based on the number line itself, as well as on a certain arbitrary assumption about how numbers occur in the real world. For once attention was given where due, albeit ever so slowly and gradually. Renewed interest in the field is increasing with an explosion in the number of published articles in recent years regarding applicability, new applications, theory, explanations, and so forth.

## BENFORD'S LAW

---

That low digits occur in the first order much more often than high digits can be seen directly from the analytical expression:

$$\text{Probability}[\text{1st digit is } d] = \text{LOG}_{10}(1 + 1/d).$$

This is known as **Benford's Law (BL)**. The implied set of proportions (Fig. 1.11) is known as the **logarithmic distribution**. Consideration is also given to second leading digits, third significant digits, and so forth, for higher orders of appearances of digits. The Generalized Benford's Law takes into account all higher-order leading digits, and it will be stated and detailed in later chapters. One remarkable consequence of BL is that the proportion assigned to digit 1 is approximately sixfold the proportion assigned to digit 9! For digit 1,  $\text{LOG}_{10}(1 + 1/1)$  or  $\text{LOG}_{10}(2)$  yields 0.301, namely 30.1%. For digit 9,  $\text{LOG}_{10}(1 + 1/9)$  or  $\text{LOG}_{10}(10/9)$  yields 0.046, namely 4.6%. Figure 1.11 shows Benford's Law for the first digits, namely  $\text{LOG}_{10}(1 + 1/d)$ . Figure 1.12 charts Benford's Law for the first digits as a histogram.

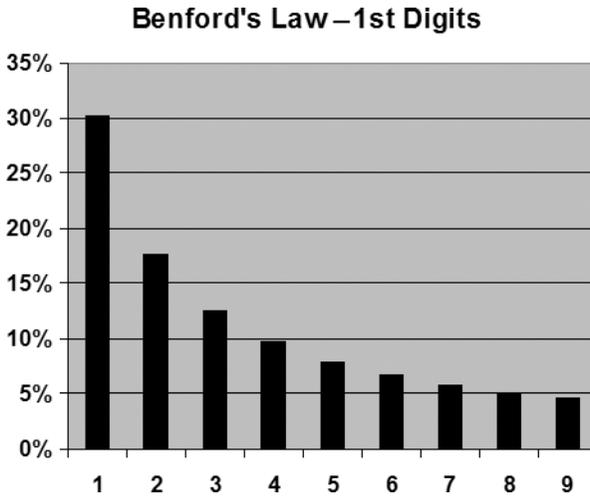
The law also describes an exact distribution for the second-order digits. For example, the second leading digit (second from the left) of 603 is digit 0, of 0.0002867 it's digit 8, and of 1,653,832 it's digit 6. It is noted that for the second and all higher orders, digit 0 is also included, and all 10 digits  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$  are in use.

613	→ digit 1
0.0002867	→ digit 8
1,653,832	→ digit 6
-0.456398	→ digit 5
603	→ digit 0

These second-order proportions among the digits are by far more equal than those of the first order. Second-order leading digits distribution (unconditional probabilities) according to Benford's Law is given in the table of Fig. 1.13, and its related chart as histogram in Fig. 1.14. Exact algebraic expressions are given in a

Digit	Probability
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

**Figure 1.11** Benford's Law — 1st Digits



**Figure 1.12** Benford's Law — Chart of 1st Leading Digits Distribution

later chapter on higher-order leading digits as they are more complex than the simple first-order algebraic expression above.

Digital proportion for the second order is not nearly as skewed in favor of low digits as is the case for the first order. Proportion for digit 0 is only 1.41 times the proportion for digit 9.

The third-order digit distribution is even more equal than for the second order. Finally there is almost total digital equality for the fourth and higher orders, where distributions are uniform and equal for all practical purposes. Figure 1.15 shows

Digit	Probability
0	12.0%
1	11.4%
2	10.9%
3	10.4%
4	10.0%
5	9.7%
6	9.3%
7	9.0%
8	8.8%
9	8.5%

Figure 1.13 Benford's Law — 2nd Digits

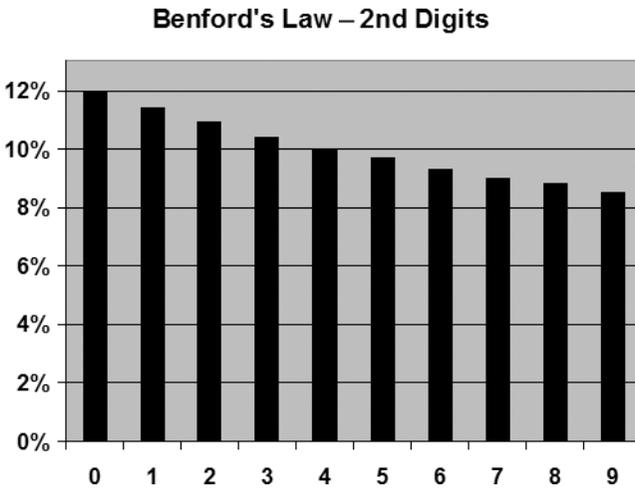


Figure 1.14 Benford's Law — Chart of 2nd Leading Digits Distribution

Benford's Law for the unconditional probabilities of the third digits. Figure 1.16 shows Benford's Law for the third digits as a histogram. Figure 1.17 depicts the chart of the distributions of the first, second, and third digital orders superimposed.

The probability of any First-Two-Digits combination (FTD), say 34 and exemplified in numbers such as 348, 0.03417, 3400.79, and so forth, is given by the expression:

$$\text{Probability [1st digit is } p \text{ AND 2nd digit is } q] = \text{LOG}_{10} (1 + 1/pq).$$

Digit	Probability
0	10.18%
1	10.14%
2	10.10%
3	10.06%
4	10.02%
5	9.98%
6	9.94%
7	9.90%
8	9.86%
9	9.83%

Figure 1.15 Benford's Law — 3rd Digits

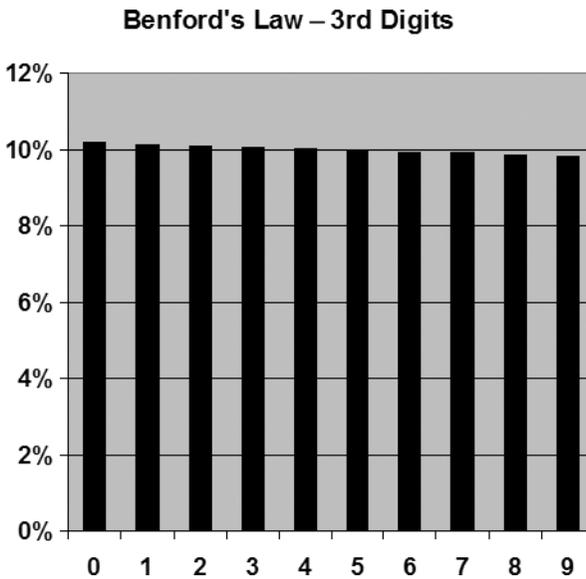


Figure 1.16 Benford's Law — Chart of 3rd Leading Digits Distribution

The term  $pq$  does not refer to the multiplication of  $p$  by  $q$  as in  $p^*q$ , but rather it refers to the creation of the number  $p^*10 + q^*1$ , where  $p \geq 1$  and  $q \geq 0$ . For example, probability of the lowest possible digital combination, namely 10, is given by  $P(10) = \text{LOG}(1 + 1/10) = \text{LOG}(1.1) = \mathbf{0.0414}$ . Also,  $P(25) = \text{LOG}(1 + 1/25) = 0.0170$ . Probability of the highest possible digital combination 99 is given by

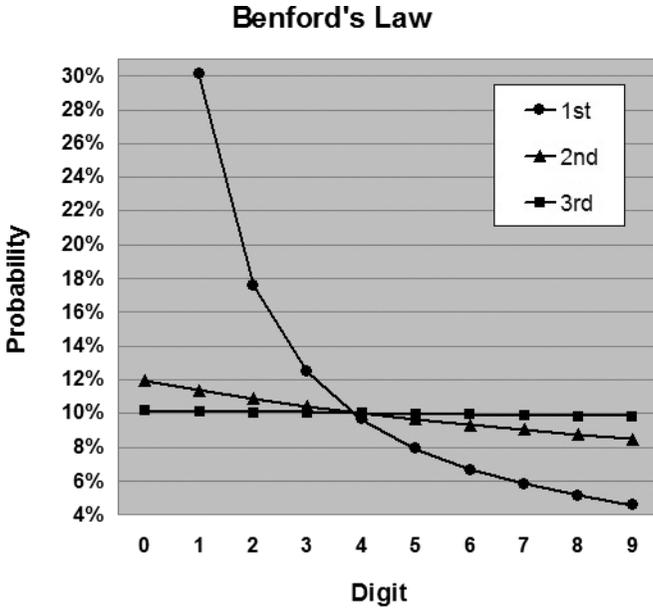


Figure 1.17 Benford's Law — Chart of 1st, 2nd, and 3rd Digits Superimposed

$P(99) = \text{LOG}(1+1/99) = \mathbf{0.0044}$ . Hence, the probability of obtaining digit combination 10 is about ten times the probability of obtaining digit combination 99! Of note is that there are 90 possibilities of the first two digital combinations, namely  $\{10, 11, 12, \dots, 97, 98, 99\}$ . The chart in Fig. 1.18 is the histogram of FTD according to Benford's Law.

The probability of any First-Three-Digits combination, say 374, and exemplified in numbers such as 37428, 0.0374317, 37400.79, and so forth, is given by the expression:

Probability [1st digit is  $p$  AND 2nd digit is  $q$  AND 3rd digit is  $r$ ] =  $\mathbf{\text{LOG}_{10}(1 + 1/pqr)}$ . The term  $pqr$  refers to the creation of the number  $p*100 + q*10 + r*1$ , where  $p \geq 1, q \geq 0$  and  $r \geq 0$ .

Typically numbers in real-life data sets are not too short (digit-wise), namely that there are normally plenty of digits within each number (say over 4 or 5 digits). In such cases, the **last** digit distribution is actually the fourth, fifth, or even some higher-order distribution, and therefore it should be about uniform with an equal probability of  $1/10$  for each digit. In fact, in such cases the last-digit distribution is normally a bit of a mixture of higher orders (fourth, fifth, etc.), but since they all

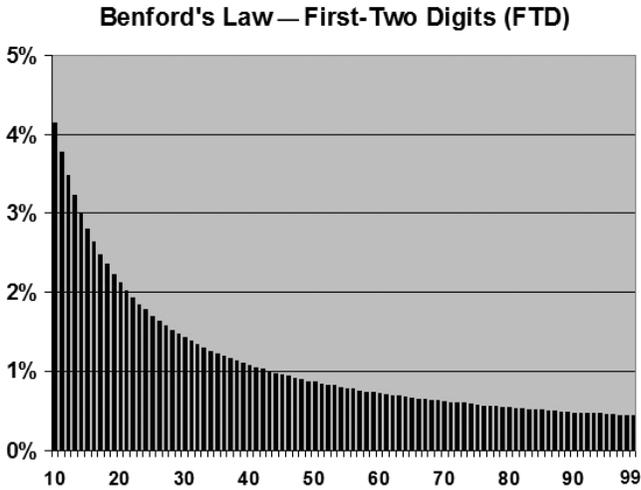


Figure 1.18 Benford's Law — Chart of First-Two Leading Digits Combination (FTD)

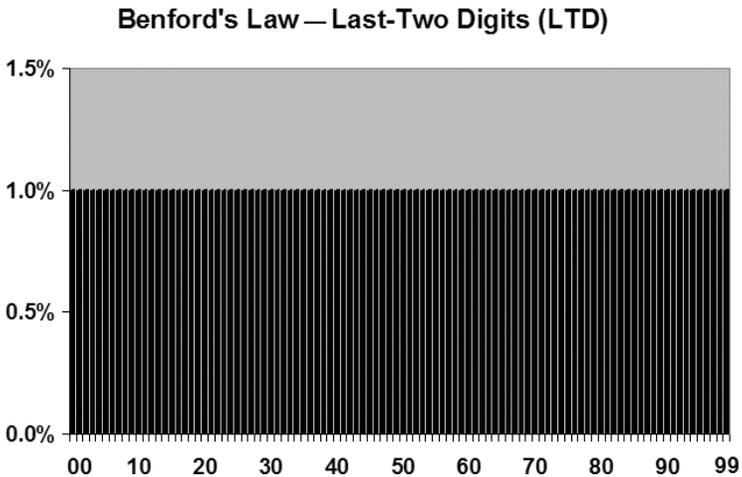


Figure 1.19 Benford's Law — Chart of Last-Two-Digits Combination (LTD)

are approximately uniform, so is the distribution of the last digit. Consideration is also given to the Last-Two-Digits combinations (LTD), which should show uniformity as well, with an equal probability of  $1/100$ , that is, 1% per pair of digit combination. This is so since there are  $10 \times 10$  or 100 possibilities of combinations here, namely  $\{00, 01, 02, \dots, 97, 98, 99\}$ . For example, probability of LTD

combination 39, exemplified in numbers such as 21739, 0.041639, 10039, and 2000.39, is simply 1%. The chart in Fig. 1.19 is the histogram of LTD according to Benford's Law.

The set of proportions of the first order in Benford's Law may seem somewhat odd, but mathematicians immediately recognize them as something quite familiar, relating directly to the logarithmic function  $\log(x)$ . The expression of the law, namely  $\text{LOG}_{10}(1+1/d)$ , utilizes a mixture of addition, division, as well as the logarithmic function, although that is not the reason for the name '**The Logarithmic Distribution**'. That name is derived from a more profound and very direct connection between the law and logarithms. One could rewrite  $\text{LOG}_{10}(1+1/d)$  as  $\text{LOG}_{10}(d/d+1/d)$  or as  $\text{LOG}_{10}((d+1)/d)$ , and using the logarithmic identity  $\text{LOG}_B(N/D) = \text{LOG}_B(N) - \text{LOG}_B(D)$ , first-digit distribution can then be expressed as  $[\text{LOG}_{10}(d + 1) - \text{LOG}_{10}(d)]$ , hence:

$$\text{Probability}[1\text{st digit is } 1] = \text{LOG}_{10}(2) - \text{LOG}_{10}(1)$$

$$\text{Probability}[1\text{st digit is } 2] = \text{LOG}_{10}(3) - \text{LOG}_{10}(2)$$

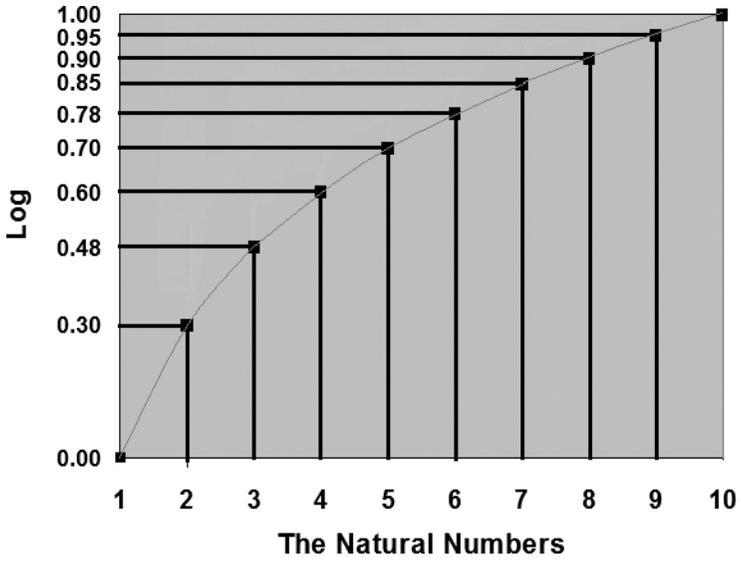
$$\text{Probability}[1\text{st digit is } 3] = \text{LOG}_{10}(4) - \text{LOG}_{10}(3)$$

In other words, the probability set

$\{30.1\%, 17.6\%, 12.5\%, 9.7\%, 7.9\%, 6.7\%, 5.8\%, 5.1\%, 4.6\%\}$  is nothing but the **differences** in the set of the logarithms of the natural numbers 1 to 10; that is, differences in  $\{\log(1), \log(2), \log(3), \log(4), \log(5), \log(6), \log(7), \log(8), \log(9), \log(10)\}$ , or differences in  $\{0.000, 0.301, 0.477, 0.602, 0.699, 0.778, 0.845, 0.903, 0.954, 1.000\}$ .

This would become all the more evident when the Generalized Benford's Law is considered in later chapters where cumulative and all other types of probabilities are expressed directly and solely by way of  $\log(x)$  function.

Figure 1.20 depicts the set of the natural numbers  $N_i$  and their associated  $\text{LOG}(N_i)$  values, enabling visualization of the logarithmic as simply the differences in values on the vertical Y axis. Note that the thin gray curve is the classic log function  $f(x) = \log(x)$ .



**Figure 1.20** Benford's Law as the Differences in Logs of the Natural Numbers

## THE PREVALENCE OF BENFORD'S LAW

---

---

Remarkably, Benford's Law is found in a bewildering variety of real-life data sets relating to physics, chemistry, astronomy, economics, finance, accounting, geology, medicine, biology, engineering, and government census data, to name just a few! The following shortlist of topics and cases is but one extremely small sample from the much larger manifestation of this ubiquitous and extraordinarily relevant law:

- Earthquake's depth below the ground
- Time interval between consecutive earthquakes
- Time interval between reversals of the Earth's geomagnetic field
- Most other geological data types
- Brightness of astronomically observed objects
- Mass of exoplanets (within our galaxy but outside the solar system)
- Distances of stars within the Milky Way galaxy to our solar system
- Rotation frequencies of spinning star remnants known as pulsars
- The molecular mass of a list of widely used chemical compounds
- Emissions of greenhouse gases per country (CO<sub>2</sub> equivalent)
- The small set of all the fundamental physical constants in physics and chemistry
- Lengths of rivers (worldwide, approximately)
- Amount of water in river flow (worldwide, almost exactly)
- Populations per county/province/district/canton/prefecture/country
- Populations centers of cities/towns/metropolitans
- Election results by district/province/city/town (if free and fair, not manipulated)
- Other census data
- General accounting data (if honest, not manipulated)
- Corporate expense/income/revenue data
- The market aggregate (combining multiple companies) of almost any item from the financial statement reports, such as inventory, total assets, shares outstanding, total number of employees, market capitalization, and so forth

- Other types of financial data
- Household income within any one country or city
- International GDP data when all countries are considered (worldwide)
- International Purchasing Power Parity (worldwide)
- Other international macroeconomic and microeconomic statistics
- Total numbers of cases by country of (almost any type of) infectious diseases reported to the World Health Organization (worldwide)
- Size of files in megabytes on any typical computer
- Global temperature anomalies 1880–2008
- Exponential growth series (almost all)
- Exponential decay series (almost all)
- The Fibonacci sequence

## PHYSICAL LAW VERSUS NUMERICAL LAW

---

---

Does Benford's Law spring from some profound physical property of nature, reflecting scientific reality? Or is it rather purely a mathematical fact, a mere consequence of our 'arbitrarily' constructed number system and digits — totally divorced from the physical world? Is it perhaps a combination of both, namely the usage of our peculiar number system to represent that peculiar physical reality as we understand and sense it in our current epoch? On the face of it, if diverse values such as 0.00347, 3.04, 35, and 3008 are all being led by digit 3 just the same, then Benford's Law does not seem to be about measurable physical values at all, but rather purely a digital law. On the other hand, for a data set having its quantities restricted to the interval (10, 100) for example, the law does refer to actual quantities, indirectly **favoring lower quantities** such as 13, 18, 22, 30, over high ones such as 67, 81, 93, 98, just as it does **favor lower digits** over high ones. If Benford's Law by extension is also a statement about physical quantities in the real world, and not merely about the symbolic aspects of numbers (digits), then more attention and prominence should be given to it in science and technology, in addition to the widespread applications that it has earned in the context of forensic data analysis for fraud detection.

It should also be emphasized that Benford's Law is a mathematical and statistical fact about how numbers are **used** and **occur** in real typical data expressing physical quantities as well as abstract entities we fancy contemplating and recording. The phenomenon is **not** purely a mathematical law about our number system itself, totally divorced from its use regarding the physical world. Had the law been purely about our number system per se, then we should have considered it a universal rule and never observe any exceptions, yet surely we do often observe exceptions to this digital rule. Newcomb phrased his statement as such and wrote: *"The law of probability of the occurrence of numbers is such that, etc."* Benford strongly believed it was a physical phenomenon, and it is quite moving to read his eloquently written final words in his article: *"As has been pointed out before, the theory of anomalous numbers is really the theory of phenomena and events, and the numbers but play the poor part of lifeless*

*symbols for living things*". Weaver misguidedly held the view that it was a "built-in characteristic of our number system", namely a purely mathematical phenomena divorced from the usage of numbers in the real world.

What was originally lacking in the formal statement of Benford's Law, and unintentionally omitted by Newcomb, Benford, and others early on, concerns a well-defined mathematical and statistical process, an exact sample space for which the law should be applied and examined. Since the most immediate concern of anyone attempting to apply Benford's Law is a clear specification of which real-life data sets exactly fall under its protective umbrella and which are outside and free to deviate, a clear guideline of targeted area of application is sorely missing. The logarithmic distribution definitely agrees with many mathematical series and algorithms defined deterministically, as well as with incredibly numerous random data types, variables, and processes, but it is certainly not a universal law true for all data types. Therefore, in order to facilitate discussion and analysis, a particular interpretation of the law will then be considered in this book and defined as follows: Benford's Law explicitly applies to the totality of all the numbers gathered from everyday and scientific data in our current modern era, a law about how numbers are being used, occur, and recorded in the physical world or in any abstract manner when considered in its absolute aggregate. In other words, the law predicts the logarithmic distribution for the significant digits of that vast collection of all the numbers printed in all the countries, in every city, in all the books and newspapers, relating to all topics, residing in all archives worldwide, and including data on all existing computers public or private. This particular interpretation will be referred to as the **Aggregate Global Data Interpretation** or **AGDI**, and that colossal data set itself simply as **AGD**. Such an interpretation though is severely limited, as it omits an incredible number of real-life data sets, data types and random processes all of which are fully Benford in their own right, as seen from the long list presented earlier, yet since such an interpretation helps in part to tell the complex and elaborate story of digit distributions it shall be frequently used in this book. Surely, that colossal collection of numbers of AGD cannot be formally defined as an exact mathematical entity, and it can never be measured directly. It is even doubtful that one could take a truly random sample from it without spending some weeks or months, least it could be claimed that data collected was too narrowly focused. As it happened, the totality of our everyday real data considered as one vast collection of numbers closely mimics a certain (abstract) random process that is itself in perfect conformity with the logarithmic law as was formally shown by Theodore Hill and shall be discussed in later chapters. This constitutes part of the motivation behind our Aggregate Global Data Interpretation of the law.

## NATURE'S WAY OF COUNTING SINGLE-ISSUE PHENOMENA

---

---

One of the most profound reasons for the prevalence of the logarithmic distribution in real-life data sets is its physical manifestation. This phenomenon springs from the widely accepted observation that Mother Nature discriminates against the large and strongly favors the small, frequently chanting her motto 'small is beautiful'. Empirical evidence consistently shows that most data sets of single-issue physical quantities are nearly perfectly logarithmic in their own right, individually considered, **provided that order of magnitude of the spread of the data is large enough**. For example, physical data falling over a narrow range of say (90, 780) is typically not logarithmic. Data falling on a wide range of say (2, 15000) is almost always logarithmic. In our digital context and for data with values over 1, 'narrow' or 'wide' ranges refer approximately to the difference between the digital sizes of the integral parts of the leftmost point and the rightmost point of the entire range. If data falls on (10, 1000) for example, then the change in digital sizes between 10 and 1000 is two digits. A simple rule of thumb requires that the difference in the logs of the edges of the range be at least 3, which means approximately a change in the digital size of at least three digits. For the hypothetical example of (90, 780), log difference is  $[\text{Log}(780) - \text{Log}(90)] = [2.89 - 1.95] = 0.94$ . For the hypothetical example of (2, 15000), log difference is  $[\text{Log}(15000) - \text{Log}(2)] = [4.18 - 0.30] = 3.88$ . This accounts for the difference in their digital behavior. Data sets having log difference of more than 3 are typically logarithmic, while data sets with small log difference are not logarithmic.

A very liberal definition of this crucial and general criterion in Benford's Law is: **Order Of Magnitude (OOM) =  $\text{Log}(\text{Max}) - \text{Log}(\text{Min}) > 3$** . Unfortunately, this definition, plus the statement above that OOM value of 3 is approximately sufficient for logarithmic behavior is too liberal, unless the definitions of maximum and minimum values exclude outliers (at a minimum) and also some data at the edges. Further analysis about OOM and logarithmic behavior shall be given in

Chapter 46 on Random Linear Combination and Revenue Data, where the more appropriate and general rule is given by:

$$\text{Order of Magnitude of Variability (OMV)} = \text{LOG}(90\text{th percentile}) - \text{LOG}(10\text{th percentile}) > 3.$$

Examples of single-issue physical manifestations of the logarithmic distribution include: amount of water in river flow, earthquake depth below the ground, time between successive earthquakes, rotation rates of pulsars, brightness of space objects, and population data, to mention just a few. Mother Nature simply counts logarithmically many or perhaps most of her phenomena. Here, one may postulate countless mini Benford's Laws: a law for earthquake depth, a law for population centers, a law for rotation rates of pulsars, and countless other mini laws referring directly to the specific physical phenomenon in question. Naturally one would seek a unified principle here, finding a single explanation for all of these diverse physical processes. Such a quest though might turn out to be elusive, as perhaps there exist more than just one driving force at play here, and that physical phenomena must be classified into two, three, or more distinct categories, each being logarithmic for different (mathematical) reasons altogether.

An instructive and forceful demonstration of the prevalence and importance of Benford's Law in the physical sciences is found in data on measurements relating to all known exoplanets in our Milky Way Galaxy, namely planets outside the solar system. As of early September 2012 there were 834 known exoplanets. The website <http://exoplanet.eu/catalog/> provides detailed data on those exoplanets, including values for planets' mass, angular distance, semi-major axis size, orbital eccentricity, and orbital period. All five data sets turned out to be in close conformity to Benford's Law! The very fact that five different aspects or measurements of the same physical reality are all nearly Benford is quite intriguing! The table in Fig. 1.21 shows the first-digit distributions of those five variables pertaining to the same single physical set of 834 planets (The % sign would be typically omitted from now on for brevity). Results show that deviations from the logarithmic are not large at all, in spite of the fact that this data set is (statistics-wise) extremely small. This current count of 834 planets represents a tiny fraction of the estimated 160 billion or so star-bound planets that exist in our galaxy. Had we had data on one billion such planets, or even 'merely' one million, it is almost a certainty that Benford's Law would have been observed almost perfectly.

Yet, with data on solely 834 planets available in our epoch, we are still able to match the Benford's proportions quite closely. This is quite significant, and it strongly

Digit	Planet's mass	Angular distance	Semimajor axis size	Orbital eccentricity	Orbital period
1	30.0	30.0	28.2	28.2	27.4
2	18.8	16.6	19.4	21.7	14.6
3	11.9	13.7	11.9	13.6	17.7
4	7.4	8.0	13.4	11.6	13.9
5	8.0	7.7	9.2	7.2	7.8
6	7.6	8.6	5.7	6.1	5.6
7	7.4	5.1	4.1	4.2	4.5
8	4.6	5.4	5.0	4.6	4.2
9	4.4	5.0	3.1	2.8	4.5

**Figure 1.21** Benford's Law is Found in Five Different Aspects of Exoplanets!

indicates that the Benford phenomenon must have deeper implications to science in general. It is noted that out of those 834 planets, only 30 were discovered in the years 1989 to 1999, while the vast majority of them, 804 in all, are very recent discoveries during the years 2000 to 2012. None was confirmed prior to 1989.

The evidence that Benford's Law is a common feature across the physical sciences spanning almost every discipline is quite compelling, and more so with the recent avalanche of additional findings and testing. Empirically, the logarithmic is found not only in data sets rooting in macrocosmic systems (stars/galaxies/rivers), but also in microcosmic systems (atomic/subatomic particles/molecules). Moreover, Benford's Law is observed in dynamic systems (earthquakes/rotations/active) as well as in static systems (molar mass/planet mass/passive). Statisticians and mathematicians will continue to debate the theoretical justification for Benford's Law, and the fact that it appears so frequently in numerous natural phenomena would not surprise them in the least, yet it does often shock many scientists!

The tables shown in Figs. 1.22 and 1.23 are courtesy of Malcolm Sambridge of the Australian National University in Canberra. The data was compiled by him and his colleagues, providing a list of natural phenomena with properties that follow Benford's Law. In the order presented in the tables, it includes: (1) the depths of almost 250,000 earthquakes that occurred worldwide between 1989 and 2009; (2) the time interval in seconds between consecutive earthquakes worldwide in the period 01/01/1970 to 12/31/2009 with no restrictions on geographical position, depth or magnitude; (3) the rotation rates or frequencies of spinning

Digit	Earthquake Depth	Time Between Earthquakes	Pulsars Rotation Frequency	River Lengths (Canada)	Global Temperature Anomalies	Global Infectious Diseases
1	31.6	29.1	33.9	21.5	27.7	33.7
2	16.9	17.2	20.7	17.1	19.4	16.7
3	14.0	12.6	12.7	14.6	12.7	13.2
4	8.7	10.0	7.6	15.8	12.1	10.7
5	7.0	8.2	5.3	9.5	8.9	7.3
6	7.4	6.9	5.0	6.3	5.4	5.4
7	5.3	5.9	4.9	6.3	6.6	4.6
8	4.6	5.2	4.7	5.1	4.3	5.1
9	4.4	4.6	4.9	3.8	2.8	3.3
Data points:	248,915	2,258,653	1,861	158	1,527	987

Figure 1.22 Earthquake, Pulsar, River, Temperature, and Disease Data that are Benford

Digit	Geomag Field	Geomagnetic Reversals	Seismic P Wavespeed (SW-Pacific)	Whole Mantle Wavespeed	Whole Earth Shear Anisotropy	Fermi Telescope Ray Fluxes
1	28.9	32.3	30.0	32.0	30.7	30.3
2	17.7	19.4	17.6	17.4	14.6	17.9
3	13.3	13.9	13.3	12.4	11.0	13.0
4	9.4	11.8	9.8	9.1	9.3	9.9
5	8.1	5.3	7.9	7.3	8.6	7.6
6	6.9	4.3	6.4	6.5	7.6	7.0
7	6.1	3.2	5.6	5.9	6.9	5.2
8	5.1	5.4	4.9	4.9	6.0	5.2
9	4.5	4.3	4.5	4.6	5.3	2.7
Data points:	36,512	93	423,776	10,000	20,544	1,451

Figure 1.23 Geological and Celestial Data that are Benford

remnants of dead stars also known as pulsars, given in Hz, from the ATNF catalogue; (4) river lengths in Canada; (5) global monthly averaged temperature anomalies from the GISTEMP database over the period 1880–2008 measured in degrees; (6) total numbers of cases of 18 infectious diseases within countries reported to the World Health Organization by 193 countries worldwide in 2007; (7) the Earth's geomagnetic field model gufm1; (8) time in years between the 93 particular reversals of the Earth's geomagnetic field; (9) regional body wave

seismic model; (10) the global seismic tomography shear wavespeed model of the Earth mantle; (11) the anisotropic shear wave mantle model saw642an; (12) the brightness of gamma rays that reach Earth as recorded by the Fermi Gamma-ray Space Telescope across the galactic in the first 11 months of operation.

Sambridge provides an interesting account of a new application to Benford's Law in his 2011 article *Benford's Law of First Digits: From Mathematical Curiosity to Change Detector*. His team of geologists was able to deduce the occurrence of an earthquake from the leading-digit distribution of available data. Besides measuring earthquake depths, Sambridge's team also examined vertical displacements of the ground in Peru as the tsunami-triggering Sumatra-Andaman earthquake of 2004 progressed. A set of ground shifts before the earthquake proper did not follow Benford's Law, but shifts that occurred during the quake itself did. The team also examined seismic data recorded at the same time by a station in Canberra. The overall patterns in the shifts persisted but the exact extent of the adherence to Benford's Law varied differently over time than in the Peruvian measurements. The team then looked more closely at Canberra seismograms and found that they were consistent with a minor, local earthquake occurring at the same time, which could be the source of the discrepancy between the two measurements. This demonstrates the ability to detect an earthquake just from the first-digit distribution of the seismic waveforms data, and in spite of the apparent loss of the complex information contained in the actual data itself – being reduced to just its first-digit proportions! “That’s the first time I know of where something physical like that was actually discovered using Benford’s Law,” said Ted Hill upon learning of the team’s work. Inspired by this remarkable example, quantum physicists are recently applying Benford’s Law to detect quantum phase transitions with success. In general, the suggested application is to first determine empirically which phenomenon actually obeys Benford’s Law, and then to deduce (or suspect) unusual processes or departures from the norm by observing any possible changes from expected digital distribution, since such digital changes are always the features of some intrinsic deviation in the state of the examined phenomenon.

## CASE STUDY I: TIME BETWEEN EARTHQUAKES

---

A case study of real-life data is given by the geological measurements of time between successive earthquakes. This data set shall be used repeatedly throughout the book demonstrating many essential aspects of Benford's Law. Three factors motivated the selection of this particular data set as a case study: (1) the relatively large data size of 19,451 values is statistically sufficient in drawing significant conclusions; (2) the near perfect logarithmic behavior of the data set; (3) the reliability, honesty, and integrity of the data provided by the U.S. Geological Survey Organization.

The data can be downloaded from the U.S. Geological Survey Organization website at <http://earthquake.usgs.gov/earthquakes/eqarchives/epic/>. The particular data set selected and obtained pertains to worldwide earthquakes, of any Richter scale magnitude, occurring at any depth below the ground, during the entire year of 2012. It is quite unsettling and worrisome that in merely one year planet Earth shook violently 19,452 times; and that is on top of our other threats and worries such as global warming, accelerating environmental degradation, wars, and a huge stockpile of nuclear and chemical weapons. On average, each 27 minutes there was an earthquake somewhere in the world! Equivalently, there were on average 53 earthquakes per day. The data set of 19,451 time intervals between earthquake occurrences created for further study utilizes the unit **second** as the scale for time.

The 1st order digit distribution is:

Time Between EQs — {29.9, 18.8, 13.5, 9.3, 7.5, 6.2, 5.8, 4.8, 4.2}

BL 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

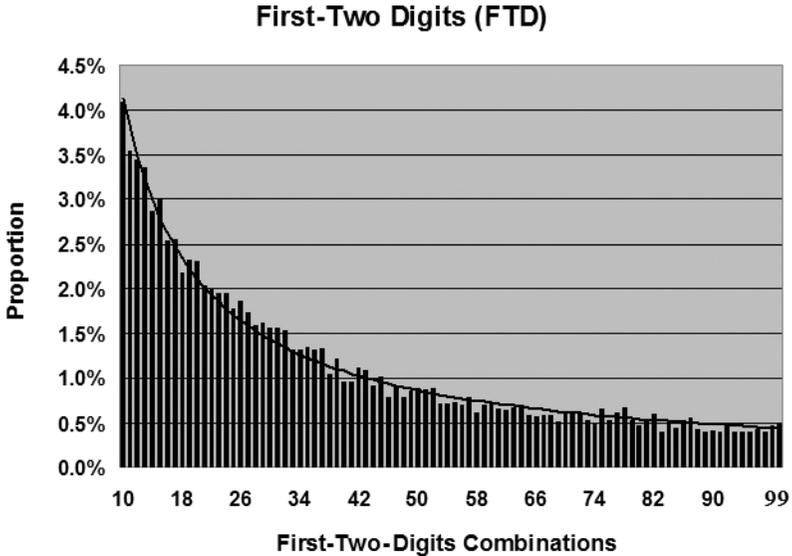
Time Between EQs — {12.0, 11.2, 11.2, 10.4, 9.8, 9.9, 9.2, 9.4, 8.3, 8.6}

BL 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

The 3rd order digit distribution is:

Time Between EQs — {10.23, 9.88, 10.15, 9.78, 10.25, 10.10, 10.10, 9.95, 9.86, 9.71}  
 BL 3rd Digits — {10.18, 10.14, 10.10, 10.06, 10.02, 9.98, 9.94, 9.90, 9.86, 9.83}

All three digital orders of the earthquake data set closely match the three distinct theoretical Benford proportions! Figure 1.24 depicts the first-two digits distribution of time between earthquake data. The smooth continuous line in the figure is the logarithmic proportion of  $\text{LOG}(1 + 1/pq)$  for FTD theoretical probabilities. Since both first-digit as well as second-digit distributions were shown above to come very close to the logarithmic separately, it is not surprising then that the combined first-two digit chart should also follow the Benford proportions very closely. Indeed the FTD histogram of the data closely matches Benford's line with very tiny and mild deviations. Figure 1.25 depicts the last-two digits distribution, where almost all combinations come out roughly between 0.5% and 1.5%, except a mild spike at 00, and a much milder spike at 50.



**Figure 1.24** First-Two Digits for Time between 2012 Earthquake Data set

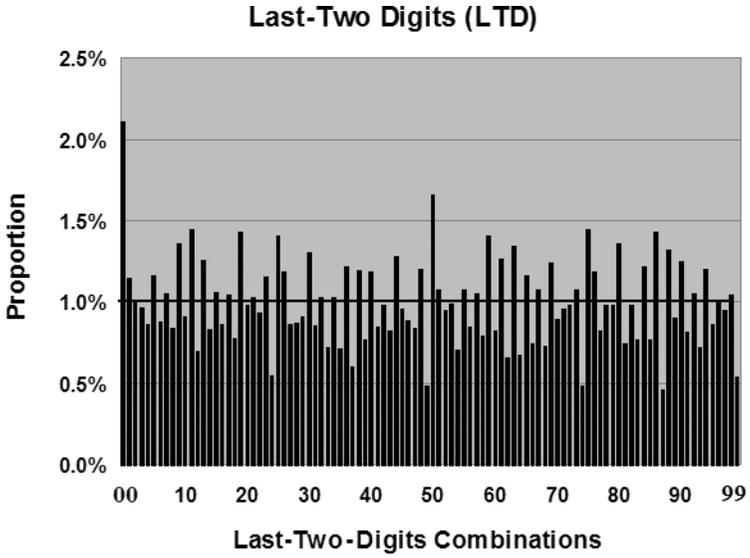


Figure 1.25 Last-Two Digits for Time between 2012 Earthquake Data set

## DATA ON POPULATION COUNTS OF CITIES, TOWNS, REGIONS, AND DISTRICTS

---

---

With the exception of a few thinkers and philosophers belonging to the deterministic school of thought such as Democritus, the much admired Baruch Spinoza, and others, most people firmly believe in free will, and that they are truly free to roam around the planet and move into cities and towns as they see fit. But Mother Nature smiles benevolently from above and arranges them all according to her own plan and quantitative design. True, she does at times let a few strong-willed individuals choose their own residences and destinies escaping that vicious deterministic cycle, but these are rare exceptions not interfering much with her overall work; for in the aggregate when it comes to humanity as a whole she has the final say on the sizes of their cities and towns, down to the smallest hamlets. She strongly prefers that they live in small or even medium-size centers, sharing their lives together. She lets very few big cities exist, but severely restricts the number of unfriendly and alienating mega metropolitans.

Indeed, data on population counts of all existing cities and towns within a given country is nearly perfectly logarithmic. Exceptions are found only for some very small countries where the size of the data is deemed to be too small in a statistical sense. The global list of population sizes of all cities and towns is extremely close to the logarithmic, due to its very large size. In addition, population counts within an artificial or arbitrary (legal) divisions or boundaries, such as counties, cantons, prefectures, regions, districts, and states (each containing multiple concentration points of cities and towns) are in principle just as perfectly logarithmic as populations centers are, except that due to their (typically) small data size, deviations from the logarithmic are usually substantial. For example, in the USA there are only 50 states, and such data size is too small to manifest its logarithmic property. On the other hand if U.S. population by county is to be considered, having about 3,143 or so counties countrywide, then strong logarithmic behavior is certainly expected (and found). By far the strongest logarithmic behavior should be found in the U.S. population data by cities and towns, encompassing over 19,000 population centers.

## CASE STUDY II: U.S. CENSUS DATA ON POPULATION CENTERS

---

Another interesting case study can be found in the USA Census Data on 2009 population counts of all incorporated cities and towns. This data set shall also be used repeatedly throughout the book demonstrating many essential aspects of Benford's Law. Three factors motivated the selection of this particular data set as a case study: (1) the relatively large data size of 19,509 values is statistically sufficient in drawing significant conclusions; (2) the near perfect logarithmic behavior of the data set; (3) the reliability, honesty, and integrity of the data provided by the U.S. Census Data Bureau.

The data can be downloaded from the U.S. Census website at: [http://www.census.gov/popest/data/historical/2000s/vintage\\_2009/datasets.html](http://www.census.gov/popest/data/historical/2000s/vintage_2009/datasets.html), with the choice of "Vintage 2009 City and Town (Incorporated Place and Minor Civil Division) Population Data sets"; the selection of "All States" as the file; and the choice of the last column titled "POP\_2009" as the data set. This data set on populations of all 19,509 incorporated cities and towns starts at value 1, namely a single person living in an officially recognized town. Its top value is that of New York City with a population of 8,391,881, which may be considered an outlier of sorts in a statistical context. One odd value of 0 was omitted from the raw data, representing perhaps a single deceased person who had lived alone in an officially recognized town (which is empty of people by now.)

This rather large data set of 19,509 such population centers adheres to Benford's Law very closely. Its first-digit, second-digit, third-digit, first-two digits, and last-two digits distributions are all in close conformity with the law.

The 1st order digit distribution is:

U.S. Pop. Centers Data — {29.4, 18.1, 12.0, 9.5, 8.0, 7.0, 6.0, 5.3, 4.6}

Benford's Law 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

U.S. Pop. Centers Data — {11.9, 11.4, 11.3, 10.5, 10.2, 9.4, 9.6, 8.7, 8.8, 8.1}

Benford's Law 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

In order to calculate third-order digit distribution for U.S. population centers, it is necessary to discard 1,092 cities and towns with fewer than 100 inhabitants. Certainly the third digit for a town with only, say, 76 people does not exist! This subtraction leaves in the calculations only  $(19509) - (1092) = (18417)$  population centers having at least 100 inhabitants. (Note: The same reasoning would entail discarding 27 cities with less than 10 inhabitants for the second-order result. Yet this tiny adjustment was not incorporated in the calculations as its effect on the result is very small.)

The 3rd order digit distribution is:

U.S. Pop. Centers Data — {10.02, 10.48, 10.36, 10.13, 9.90, 10.15, 9.94, 9.69, 9.61, 9.71}  
 Benford's Law 3rd Digits — {10.18, 10.14, 10.10, 10.06, 10.02, 9.98, 9.94, 9.90, 9.86, 9.83}

All three digital orders of the U.S. population data set closely match the three distinct theoretical Benford proportions! Figure 1.26 depicts the first-two digits distribution of U.S. population centers data with its almost perfect conformity to the law, as expected since first and second digits were shown above to come very close to the logarithmic themselves.

Figure 1.27 depicts the last-two digits distribution of only a portion of the data set on U.S. population centers. From a total of 19,509 centers, 9,294 centers

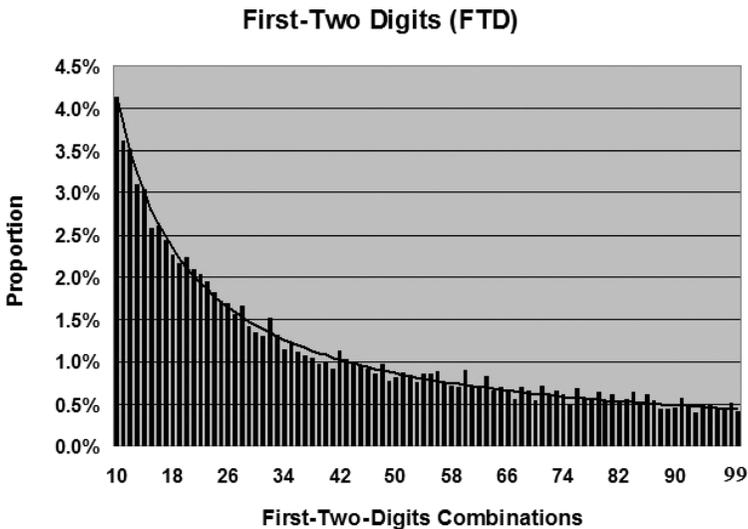
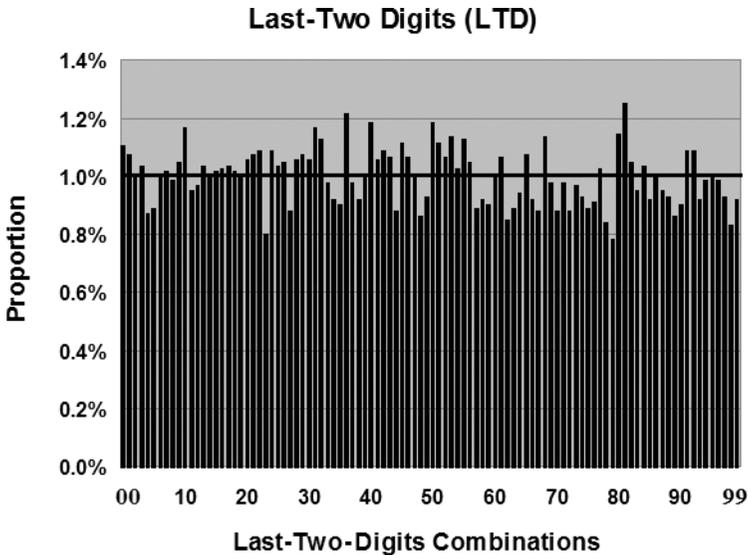


Figure 1.26 First-Two Digits for 2009 U.S. Population of Cities and Towns Data set



**Figure 1.27** Last-Two Digits for 2009 U.S. Population of Cities and Towns Data set

having population less than 1,000 people were omitted, leaving only 10,215 large population centers with 1,000 or more inhabitants to contribute to the LTD chart. The reason a small town with, say, a population of 13 inhabitants should not be included in the LTD distribution is that its chance of occurrence follows the first-two digits skewed distribution of  $\text{LOG}(1 + 1/13)$  or 3.2% rather than the last-two digits equal distribution of 1.0%. A city with only, say, 538 people should not contribute 38 to LTD, but rather should contribute 53 to FTD. A detailed discussion on digit-anemic ('small') numbers in forensic digital analysis shall be given in a later chapter. Surely the mild deviations in LTD seen in Fig. 1.27 do not appear to represent any significant violation of the law, although an exact numerical measure calculated from the vector of LTD proportions would better serve us in formally deciding on compliance. Discussion about such compliance measures shall be given in the section on data compliance tests. Admittedly deviations in LTD from the Benford line are visually slightly larger than those in FTD chart. Appearances are deceiving and one should not be misguided by the visually larger deviations in the LTD chart as compared with the FTD chart, since this perception is mostly derived from the more refined 0.2%-scale in Fig. 1.27 than the relatively cruder 0.5%-scale in Fig. 1.26. Yet if both are viewed superimposed on the same-size scale, LTD does indeed appear to have slightly larger deviations than those in

FTD. Such discrepancy may be explained away when error in recording is taken into account, which is much more prevalent in LTD than in FTD. For example, if the population count in New York City is 8,391,881, then placing 83 in FTD is certainly error-free, while placing 81 in LTD is potentially erroneous. This is due to the census office's method of collecting data, which could easily miss out on say 124 people, or wrongly overestimate the New York population by, say, 55 inhabitants, but the figure of 8.3 million is exact! It should be noted that revenue data though has by far more equal integrity on both digital ends. For example, a bill of \$34,993.50 gives almost equal credence to 34 on the FTD chart as it does to 50 on the LTD chart, even though the person writing the bill is less likely to err with the 34 thousands than with the 50 cents. The person or corporation having to pay such enormous amount of money would surely let an error with the 50 cents pass unnoticed, but would scrutinize the value of 34 thousand dollars very carefully before payment.

## DATA SETS ON USA POPULATION BY STATE AND BY COUNTY

---

In comparison to the large data size of population of all cities and towns in the USA with its nearly perfect logarithmic behavior, the much shorter data set on U.S. population among its 50 states (plus one count of the population within the District of Columbia) does not show such exact digital behavior. This can be studied by downloading data pertaining to the 2012 U.S. population by state from <http://www.infoplease.com/ipa/A0004986.html>. The data set of 51 points should certainly be considered as too small in a statistical sense.

The 1st order digit distribution is:

U.S. Pop. By State — {25.5, 13.7, 9.8, 7.8, 9.8, 17.6, 2.0, 5.9, 7.8}

Benford's Law 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

U.S. Pop. By State — { 9.8, 7.8, 5.9, 15.7, 2.0, 7.8, 5.9, 9.8, 23.5, 11.8}

Benford's Law 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

Yet some resemblance to the logarithmic is clearly seen here. Surely the population value of each state can be thought of as the sum of numerous population counts of all its cities and towns; ignoring the few isolated inhabitants who reside outside any officially recognized population centers.

Data on U.S. population for all its 3,143 counties nationwide is certainly of a sufficient size in a statistical sense, and therefore should be nearly logarithmic. Such data can also be found on the U.S. Census website via the link <http://www.census.gov/popest/data/counties/totals/2012/CO-EST2012-01.html>.

At the bottom of the webpage, the choice of "All States" is selected. The choice of population census as of April 1, 2010 of the first column on the left is then selected.

The 1st order digit distribution is:

U.S. Pop. by County — {30.3, 18.9, 11.9, 9.8, 6.8, 6.7, 5.8, 4.8, 5.0}

Benford's Law 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

U.S. Pop. by County — {11.9, 10.8, 10.8, 11.4, 10.2, 9.4, 9.7, 8.8, 9.1, 7.9}

Benford's Law 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

## FOUR DISTINCT NUMERICAL PROCESSES LEADING TO BENFORD

---

---

Thus far we have examined various real-life data sets and found their digits to be quite close to the logarithmic distribution. Actual data examined include: stock market volume and prices, multiplication table, address data, five aspects of exoplanets, other geological and astronomical data sets, as well as detailed examination and analysis on population count and earthquake time data. We shall now turn our attention to four abstract numerical and statistical processes that lead to the logarithmic distribution. These four processes may be thought of as causes or explanations of the whole leading digits Benford phenomenon. Remarkably, these four distinct processes all point to the same resultant digital signature, namely the logarithmic signature of  $\text{LOG}(1 + 1/d)$ ! The four processes are: (I) random linear combinations, (II) data sets aggregation, (III) random mix of numbers, and (IV) multiplication processes. These four numerical processes constitute perhaps the most important and relevant causes leading to Benford's Law, both in a theoretical sense as well as in applications. Even more remarkable is the fact that there are still other numerical processes in addition to the four mentioned in this section which also lead to the logarithmic. Account of the other processes shall be given in the fourth and fifth sections.

## RANDOM LINEAR COMBINATIONS AND ACCOUNTING REVENUE DATA

---

An important cause for the prevalence of the logarithmic distribution in numerous real-life data sets arises from one particular leading digits theoretical result relating to statistical processes generated by Random Linear Combinations (RLC) of values. The most typical and well-known such process is accounting data relating to a company's revenue amounts, namely the list of actual bills paid by customers. A typical purchase by a client shopping at an IT store might consist of one computer costing \$860, two USB keys at \$15 each, and 10 packages of CD disks at \$5 a package. The total bill of \$940 for all these products is simply derived from the linear combination of the various prices listed at the shop, namely  $1 * (\$860) + 2 * (\$15) + 10 * (\$5)$ . At play here are simultaneous random variables and decisions that determine the very limited selection of products (typically only 1, 2, or 3) from that very long price list (typically in the hundreds or thousands) relating to all available products on sale, and with the possibility of buying multiple number of units of each product chosen. That revenue data follows Benford's Law quite closely is a well-known empirical fact (given that large enough number of entries are available). Yet, revenue data is not the only real-life data types that can be modeled on random linear combinations of numbers, rather there are numerous other types that are intrinsically such, although this may not be immediately obvious when one contemplates such physical or abstract data types. The molar mass of all commonly used and synthesized chemical compounds worldwide constitutes another manifestation of RLC and therefore it is nearly logarithmic. The molar mass of a molecule is analogous to a random pick from the weights of the elements in the Periodic Table. For example, alcohol  $C_2H_6O$  has the molar mass of 46.069, and the value of its mass is derived from the combination of  $2 * (12.011) + 6 * (1.008) + 1 * (15.999) = 46.069$ . In this context, the Periodic Table serves as the price list of all the items on sale in the shop, and the molar mass serves as the total bill paid by the customer.

Total bill paid by the customer has three different levels of randomness or uncertainty:

- How many distinct items (products) the shopper will pick?
- Which items (products) will be chosen?
- How many units (quantity) will be purchased for each item (product) chosen?

The backdrop of all these random factors which strongly affects and drives the result is the price list itself, which in a sense can also be considered as a (discrete) random variable. It should be noted that the price list is typically not as in the uniform discrete distribution, since it does not normally increase (say) \$1 from one product to the next more expensive product. Rather there are typically numerous cheap items, some intermediate ones, and very few extremely expensive items for sale.

The following analysis is an example of a particular computer simulation result relating to random linear combinations of one hypothetical very small shop that has only nine items for sale. All nine items are assumed here equally likely to be sold (equal popularity). The shopper is assumed to purchase exactly two (distinct or identical) items, and this assumption removes one level of randomness in the process. The shopper is then to roll two dice (one for each item) as a way to decide on the number of units (quantities) he or she may wish to buy — out of {1, 2, 3, 4, 5, 6} possibilities. At most, when both dice show side six, the shopper will buy 12 units in total. At a minimum, when both dice show side one, only two units would be purchased (one unit per item).

The list of prices for the nine items for sale in this hypothetical shop is:

List = {\$2.25, \$3.25, \$4.75, \$7.75, \$9.50, \$10.25, \$25.00, \$35.00, \$37.00}

The expression of the schematic arrangement of these computer simulations is:

Random Linear Combination = List\*dice1 + List\*dice2

Figure 1.28 depicts a small section from the random output of these computer simulations, showing 17 hypothetical purchases. Each row represents a single simulated result. For example, the first row indicates that the shopper bought six units of the item costing \$37.00, and in addition bought six units of the item costing \$35.00, paying a total of \$432.00 for those 12 items. Monte Carlo computer simulation — invented in the late 1940s by Stanislaw Ulam while working on thermonuclear weapons projects at the Los Alamos National

Price for item	Quantity	Sub-total	Price for item	Quantity	Sub-total	Final bill
\$ 37.00	6	\$ 222.00	\$ 35.00	6	\$ 210.00	\$ 432.00
\$ 4.75	3	\$ 14.25	\$ 2.25	2	\$ 4.50	\$ 18.75
\$ 7.75	4	\$ 31.00	\$ 10.25	5	\$ 51.25	\$ 82.25
\$ 25.00	3	\$ 75.00	\$ 10.25	5	\$ 51.25	\$ 126.25
\$ 10.25	5	\$ 51.25	\$ 35.00	4	\$ 140.00	\$ 191.25
\$ 9.50	2	\$ 19.00	\$ 37.00	4	\$ 148.00	\$ 167.00
\$ 9.50	6	\$ 57.00	\$ 2.25	2	\$ 4.50	\$ 61.50
\$ 4.75	4	\$ 19.00	\$ 9.50	1	\$ 9.50	\$ 28.50
\$ 10.25	6	\$ 61.50	\$ 37.00	2	\$ 74.00	\$ 135.50
\$ 3.25	4	\$ 13.00	\$ 2.25	2	\$ 4.50	\$ 17.50
\$ 3.25	3	\$ 9.75	\$ 10.25	2	\$ 20.50	\$ 30.25
\$ 9.50	4	\$ 38.00	\$ 7.75	1	\$ 7.75	\$ 45.75
\$ 35.00	2	\$ 70.00	\$ 7.75	1	\$ 7.75	\$ 77.75
\$ 4.75	3	\$ 14.25	\$ 4.75	5	\$ 23.75	\$ 38.00
\$ 25.00	5	\$ 125.00	\$ 7.75	3	\$ 23.25	\$ 148.25
\$ 25.00	3	\$ 75.00	\$ 25.00	5	\$ 125.00	\$ 200.00
\$ 10.25	1	\$ 10.25	\$ 3.25	3	\$ 9.75	\$ 20.00

Figure 1.28 Random Linear Combinations of List of Prices Modeling Revenue Data

Laboratory — would serve as an immensely useful tool in this book in solving numerous problems and facilitating many predictions. In this case, the computer is throwing a virtual dice, giving each of the six faces equal chance of occurring. The computer also chooses randomly an item from the price list with equal probability for all the nine prices. The rest of the calculations and output are straightforward.

The focus here is on the last column on the right labeled ‘Final Bill’. Even for this tiny sample of 17 purchases, proportion of digit 1 is a whopping 7/17 or 41.2%. Proportion of digit 2 is 3/17 or 17.6%, while digit 9 never occurs in the first order. Surely, a sample of just 17 simulations is too small in a statistical sense. Five distinct batches of 2,000 simulations each are performed; the results are as follow:

- Simulation A: {30.8, 19.4, 11.3, 8.9, 8.7, 6.4, 6.3, 5.3, 3.2}
- Simulation B: {31.2, 20.1, 11.0, 8.5, 7.8, 6.8, 6.5, 4.8, 3.5}
- Simulation C: {30.2, 21.3, 10.2, 8.4, 9.3, 6.1, 5.5, 5.7, 3.5}
- Simulation D: {31.5, 20.9, 10.7, 8.0, 8.2, 7.0, 5.9, 4.6, 3.2}
- Simulation E: {31.8, 20.8, 10.5, 8.3, 7.7, 6.4, 6.4, 5.5, 2.8}
- Benford: {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}**

Remarkably, even though neither the list of prices nor the set  $\{1, 2, 3, 4, 5, 6\}$  of all possible dice faces have any bias towards low digits such as 1 or 2, nonetheless this particular process of RLC yields digital results that strongly favor low digits over high ones, and resembles the logarithmic a great deal! Only the price of \$10.25 in the list is being led by digit 1, implying  $1/9$  proportion, or 11%; yet in the simulated revenue data digit 1 leads by about 31%! It is as if **Benford's Law appears out of nowhere!**

It is very important to notice that this particular process of revenue streams from this shop is non-logarithmic as it does not follow Benford's Law exactly, yet it has its own consistent **unique leading digit (hidden) signature**. This signature may be approximately deduced from the average of the above five batches, or more exactly by simulating one large batch of say seven million purchases. Another possibility is to attempt to deduce its digital signature in a mathematical way from the setup of the numerical process itself, and to expect such results to be extremely close to the results coming out of computer simulations. Knowledge of Benford's Law could give tax authorities an ability to detect fraud whenever the shop's accountant invents fake revenue data with slightly lower income. Detecting such fraud could be done by simply examining the digital signature of provided data and comparing it to the theoretical expectation (Benford or otherwise). If digital discrepancy is small it is accepted as being merely random in nature, but when the discrepancy is large it causes the authorities to suspect fraud in reporting, and merits further investigation. The principal tool in forensic digital analysis can certainly go far beyond the confines of Benford's Law, and it proves immensely useful in fraud detection. A comprehensive analysis of RLC and detailed applications in the context of revenue data is given in the third section.

## AGGREGATION OF DATA SETS AS A PROMINENT CAUSE OF BENFORD'S LAW

---

Appending numerous data sets as one large piece of data leads to strong logarithmic behavior, provided all data sets unite in starting uniformly from around 0, 1, or another low value, but differ in the way they terminate, with some falling on short ranges while others falling on longer ones. To illustrate such logarithmic convergence with one concrete case, the following six hypothetical data sets shall be combined into a singular data set:

Data set A: {2.7, 3.1, 5.5}

Data set B: {1.6, 4.7, 4.9, 6.0, 8.5}

Data set C: {3.5, 3.9, 6.3, 7.0, 8.9, 9.8, 12.3}

Data set D: {1.4, 4.2, 7.1, 7.5, 12.5, 16.8, 19.9, 20.4, 27.3}

Data set E: {2.8, 4.8, 5.4, 5.9, 6.6, 13.3, 17.5, 27.7, 29.0, 33.3, 36.8}

Data set F: {1.8, 3.7, 8.8, 9.0, 11.1, 16.4, 17.9, 18.8, 23.3, 29.1, 39.2, 45.8, 75.5}

Separately, each of the six data sets surely is not even remotely logarithmic, yet when merged into a single data set, digital configuration is nearly logarithmic.

Data Set A-F: {1.4, 1.6, 1.8, 2.7, 2.8, 3.1, 3.5, 3.7, 3.9, 4.2, 4.7, 4.8, 4.9, 5.4, 5.5, 5.9, 6.0, 6.3, 6.6, 7.0, 7.1, 7.5, 8.5, 8.8, 8.9, 9.0, 9.8, 11.1, 12.3, 12.5, 13.3, 16.4, 16.8, 17.5, 17.9, 18.8, 19.9, 20.4, 23.3, 27.3, 27.7, 29.0, 29.1, 33.3, 36.8, 39.2, 45.8, 75.5}

Calculating 1st digits distribution for the combined data set above we get:

Data set A-F : {27.1, 16.7, 14.6, 10.4, 6.3, 6.3, 8.3, 6.3, 4.2}

Benford's Law: {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

**Remarkably, Benford's Law seems to appear out of nowhere!**

The driving force behind such logarithmic convergence is the partial overlapping of ranges in the aggregate data set, which occurs mostly on the left for low values, and which gradually diminishes on the right for high values. This

process piles up more values and digits on the left in relation to the right, resulting in skewed and uneven outcome, just as the logarithmic itself is skewed and uneven, preferring low digits which naturally are 'on the left', and discriminating against high digits which naturally are 'on the right'.

Clearly, not all aggregations lead to such piling up of values on the left. As a counter example, combining data sets which do not overlap at all, but rather continuously expand such as in  $\{1.8, 4.7, 5.8\}$ ,  $\{9.7, 13.3, 25.2\}$ ,  $\{28.9, 38.5, 44.2\}$ ,  $\{50.3, 78.7, 102.6, 577.9\}$  does not result in convergence to the logarithmic at all. Another counter example is given by the aggregation of data sets all overlapping over almost the same interval (the other extreme), and all united almost in where they start and in where they terminate, such as in the example  $\{1.5, 5.9, 9.3\}$ ,  $\{2.1, 3.8, 7.7\}$ ,  $\{2.4, 4.3, 8.6\}$ ,  $\{1.1, 4.2, 7.9\}$ ,  $\{3.6, 5.2, 7.5\}$ . Aggregating such data sets does not lead to any convergence to the logarithmic at all.

It may not be quite apparent or obvious, but this numerical process of (skewed) data aggregation actually occurs quite frequently in real-life data sets. Some such data types can be implicitly modeled on such aggregation; for other data types this is clearly so in an explicit way as narrated above. As for just one example, house numbers in street address data can be modeled on exactly such aggregation. The address list of all house numbers in a particular post office branch in a very small town somewhere in, say, Pennsylvania, USA may typically read as follows:

$\{1,2,3,4,5,6,7,8,9,10,11,12,13,14\}$  — for Maple Street

$\{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18\}$  — for Cherry Street

$\{1,2,3,4,5,6\}$  — for Allen Street, the shortest street in town

$\{1,3,4,5,6,7,8,9,10\}$  — West Street, where house # 2 was demolished last year

$\{1,2,3,4,5,6, \dots, 58, 59, 60\}$  — for Main Street, the longest street in town

While not exactly logarithmic, address data is quite close to the logarithmic distribution. This explains why house numbers in South Dakota address data seen in Fig. 1.6 came a bit close to the logarithmic, despite its very small data size.

## RANDOM PICK FROM A VARIETY OF DATA SOURCES IS LOGARITHMIC

---

Another cause of logarithmic behavior in some particular data sets is the consistent empirical evidence (backed by rigorous mathematics) that random selections from a large variety of data sources obey Benford's Law in the limit as the size of such data set gets very large. The numbers selected should be of positive values exclusively, without any negative values mixed in. Strictly speaking the process should pick only one number from a given source, then another number from another (related or totally unrelated) source, and so forth. The end result of the whole process is a large mixture of unrelated numbers, representing perhaps a meaningless data set not conveying any specific information, yet having a definite (and logarithmic) digital signature.

In order to empirically test this convergence, 90 numbers were picked from 90 different Internet pages, newspaper pages, economics-related data publications, sport data, medical data, and U.S. Census publications. No more than one positive number was picked from any given source. Figure 1.29 depicts those 90 numbers. As evident by the very diverse ranges (order of magnitude) of those 90 numbers, they arise from very different and totally unrelated data sources.

Calculating 1st digit distribution for this mixture of numbers we get:

Random Pick 1st Digits	— {24.4, 17.8, 12.2, 10.0, 11.1, 5.6, 8.9, 5.6, 4.4}
Benford's Law 1st Order	— {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

Digital configuration is quite near the logarithmic. It is not sufficiently close to it due to the small size of the data set of 90. A number collection of a few hundred such numbers would bring digits much closer to the logarithmic, and a few thousands would show a near perfect logarithmic behavior. **Remarkably, this third result once again demonstrates how Benford's Law seems to appear totally out of nowhere!**

Although the last two processes mentioned here, namely (I) random pick from a variety of sources in this chapter, and (II) data sets aggregation of the previous

4708	168	1.2	665831857	18.25
0.937182	20482	17	13	51
1989	7083	205305	2.337	375
2862	158889	7	13.56	110
120000	14.55	1537	74.46	9728
3342	439	2.16	4.532	157
0.734	65679	13064	91605	51369
5272	3.14015	2	3	11.1
1202181	255	59954	926845	82082
0.673	34.6	31	5312	26.1
300.785	42.8	4	2670773	5812.5
85	1770	3114	503724	43460
0.4	6710	710	3000586	504359.93
33.16	3505	188	25.38	7810.72
526112	0.79	51104	8722245	13
85.669	21013	0.8	28.4	229742
242	11495.53	60	2109	4249.874
44	26884	11.1231	108.9	77220

**Figure 1.29** Random Pick of 90 Numbers from 90 Unrelated Sources

chapter, may appear identical, they are not, and in one sense the difference between them can be described as in the relationship between the numbers that are being either picked or aggregated. (I) Random pick is the gathering of totally unrelated numbers, while (II) data sets aggregation is the well-structured combination of related numbers pertaining to the same phenomenon or measurement, and conveying some specific statistical message.

## INTEGRAL POWERS OF TEN

---

---

The abbreviation **IPOT** will stand for an ‘Integral Power Of Ten’ number (singular), namely  $10^{\text{integer}}$  where the integer may be negative or zero. Examples of IPOT numbers are 0.01, 0.1, 1, 10, 100, and so on, derived from  $10^{-2}$ ,  $10^{-1}$ ,  $10^0$ ,  $10^1$ ,  $10^2$ . Adjacent integral powers of ten are two sequential (neighboring) IPOT numbers, namely the pair  $10^{\text{integer}}$  &  $10^{\text{integer} + 1}$  such as 10 & 100, 1 & 10, and so forth. IPOT numbers and adjacent IPOT pairs play a crucial role in the understanding of Benford’s Law and Leading Digits in general.

Beware of calling 0 and 1 adjacent IPOT numbers; they are not! These two numbers are not adjacent in that sense, rather they are infinitely far away from each other in the digital sense. Calling attention to these two very important numbers demonstrates the unique nature of the interval (0, 1) in Leading Digits, since within that ‘short’ interval infinitely many IPOT numbers quietly reside! For example, IPOT numbers such as 0.000001, 0.01, 0.1, and  $10^{-17}$  all reside within (0, 1), and LOG on (0, 1) varies as in  $(-\infty, 0)$ .

In contrast, on the interval [1, 1000] there exist only four IPOT numbers, namely 1, 10, 100, and 1000, and LOG on this interval varies only from 0 to 3. It should be noted that 10 was chosen for being the **base** in our number system, and that all this generalizes to other bases.

## THE LOGARITHMIC AS REPEATED MULTIPLICATIONS

---

---

Of the many aspects, manifestations, and causes of Benford's Law, the case of multiplication processes stands out as one of the most important aspects of the phenomena, and a thorough understanding of it is crucial in the field. Repeated multiplication processes, where all intermediate products are considered and included, effectively drive numbers toward the logarithmic distribution in the limit. It should be noted that some multiplication processes are deterministic in nature, while others are random.

Consider the arbitrarily chosen value of 8 being repeatedly multiplied (13 times) by another arbitrary number 3; in symbols:  $8 \cdot 3^N$  where  $N = \{0, 1, 2, 3, \text{ and so forth}\}$ . Equivalently:  $\{8, 8 \cdot 3, 8 \cdot 3 \cdot 3, 8 \cdot 3 \cdot 3 \cdot 3, 8 \cdot 3 \cdot 3 \cdot 3 \cdot 3, 8 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3, \text{ and so forth}\}$ . Note that in terms of growth, this series starts out from the base of 8, and explodes upwards, tripling with each new element; its high growth rate stands at 200%. For all exponential growth series, the multiplicative factor  $F$  and the growth rate  $P$  are related via the expression  $F = 1 + (P/100)$ . For 200% growth rate for example,  $F = 1 + (200/100) = 1 + 2 = 3$ .

Figure 1.30 depicts the process and its first digits. Distribution of the 1st significant digits is:  $\{35.7, 14.3, 0.0, 14.3, 14.3, 7.1, 7.1, 7.1, 0.0\}$ . For example, there are five elements beginning with digit 1 out of 14, hence  $5/14 = 0.357$  or 35.7%.

If one considers a longer sequence of  $8 \cdot 3^N$  having 26 elements, instead of just 14, then 1st digit distribution gets closer to the logarithmic and it is almost monotonically decreasing. A slightly longer sequence having 45 elements brings us even closer to the logarithmic, and so forth. Finally, when a long sequence of

$8 \cdot 3^N$	$8 \cdot 3^N$ series	1st digit
$8 \cdot 3^0$	8	8
$8 \cdot 3^1$	24	2
$8 \cdot 3^2$	72	7
$8 \cdot 3^3$	216	2
$8 \cdot 3^4$	648	6
$8 \cdot 3^5$	1944	1
$8 \cdot 3^6$	5832	5
$8 \cdot 3^7$	17496	1
$8 \cdot 3^8$	52488	5
$8 \cdot 3^9$	157464	1
$8 \cdot 3^{10}$	472392	4
$8 \cdot 3^{11}$	1417176	1
$8 \cdot 3^{12}$	4251528	4
$8 \cdot 3^{13}$	12754584	1

Figure 1.30 Multiplication Process as in  $8 \cdot 3^N$  Series

634 elements is considered, the closeness to Benford is quite striking! Here are five sequences of  $8 \cdot 3^N$  and their 1st leading digits distributions:

- 14 elements : {35.7, 14.3, 0.0, 14.3, 14.3, 7.1, 7.1, 7.1, 0.0}
- 26 elements : {26.9, 19.2, 11.5, 7.7, 7.7, 7.7, 7.7, 7.7, 3.8}
- 45 elements : {28.9, 17.8, 11.1, 11.1, 6.7, 6.7, 6.7, 6.7, 4.4}
- 254 elements: {29.9, 17.7, 12.2, 9.8, 7.9, 7.5, 5.9, 4.7, 4.3}
- 634 elements: {30.1, 17.5, 12.6, 9.5, 8.0, 6.8, 5.8, 5.0, 4.6}
- Benford:** {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

Remarkably, once again Benford’s Law seems to appear out of nowhere! Other geometric progressions with different arbitrarily chosen values for the base and the mutiplicative factor give very similar results, so that the above particular result with base 8 and factor 3 is quite general and representative, except for some very rare anomalous cases to be discussed in the sixth section. Yet, for series with low growth, there is a second requirement (besides length considerations) for conformity to Benford’s Law relating to the first and the last elements of the series. If ones writes the expression **Last = Constant\*First**, then that Constant should be as close as possible to an IPOT number such as 10, 100, 1000 and so forth, in order to obtain the logarithmic. This second requirement for low-growth series can be equivalently stated as follow: the value of the exponent difference

between the first and the last elements should be ideally as close to an integer as possible. For example, series having their edge values (the first and last elements) as in  $\{10, 100\}$ ,  $\{1, 1000\}$ ,  $\{25, 2500\}$  perfectly fulfill this requirement, since they could be written as  $\{10^1, 10^2\}$ ,  $\{10^0, 10^3\}$ ,  $\{10^{1.398}, 10^{3.398}\}$  and exponent differences are 1, 3, and 2, respectively, and thus integers. For a vivid example of this second requirement, let us consider an exponential growth series from base 10, growing at the very low and slow rate of 3% per period; having a multiplicative factor of 1.03; namely the sequence  $10 * 1.03^N$  where  $N = \{0, 1, 2, 3, \text{and so forth}\}$ . Conformity to the logarithmic depends on the value of its ending element (the last) relative to the initial element of 10 (the first); being nearly logarithmic whenever the last element is approximately close to being IPOT multiple of the first element 10. In symbols: whenever  $\text{Last} = \text{IPOT} * \text{First}$ , namely  $\text{Last} = \text{IPOT} * 10$ , or simply  $\text{Last} = \text{IPOT}$ , since 10 itself is an IPOT and therefore can be 'absorbed' within the term IPOT. Six possible lengths will be considered, showing first digit configuration for each length, with the last sequential value of  $10 * 1.03^N$  shown in bold font on the right side of the distributions:

49 elements :  $\{49.0, 28.6, 18.4, 4.1, 0.0, 0.0, 0.0, 0.0, 0.0\}$  last = **41.3**  
 79 elements :  $\{31.6, 17.7, 11.4, 10.1, 7.6, 6.3, 6.3, 5.1, 3.8\}$  last = **100.3**  
 132 elements:  $\{36.4, 21.2, 13.6, 11.4, 4.5, 3.8, 3.8, 3.0, 2.3\}$  last = **480.5**  
 157 elements:  $\{31.2, 17.8, 11.5, 10.2, 7.6, 6.4, 6.4, 5.1, 3.8\}$  last = **1006.0**  
 208 elements:  $\{34.6, 19.7, 13.5, 10.1, 5.8, 4.8, 4.8, 3.8, 2.9\}$  last = **4542.6**  
 391 elements:  $\{30.9, 17.1, 12.3, 10.2, 7.7, 6.4, 6.1, 5.1, 4.1\}$  last = **1015120.6**

Note the oscillations of the distributions in relation to the logarithmic. Whenever the last term is near an integral power of ten (i.e. exponent difference is close to being an integer) digit distribution closely follows the logarithmic, but whenever its last value is far from such a value it does not follow the logarithmic. In the limit, as the length of the series gets extremely large (in terms of passing numerous IPOT values), this second requirement is waived, since those fluctuations seen above always die out eventually, and digits converge to the logarithmic. The earlier example of the sequence  $8 * 3^N$  when considered as an exponential growth series represents unusually high growth of 200% per period, and such an exceedingly fast growth frequently passes an IPOT number almost every other sequence, rendering the second requirement unimportant. To grow by 200% implies a multiplicative factor of 3.0, and in two periods it grows by  $3 * 3$  or 9, namely by 800%, therefore approximately it is almost being multiplied by the number 10 each two periods, guaranteeing a passage by an IPOT number roughly

every two sequences. Why does it matter where the series starts or ends for logarithmic behavior? Why is it usually necessary to have a long sequence for logarithmic convergence? Why does passing through many IPOT numbers bring about convergence? The general understanding of all the factors and machinations at play here will become apparent in later chapters in the mathematical section; only a basic sketch of the issue is presented in this chapter.

In any case, why should this particular abstract property of numbers under constant arithmetical attacks have anything to do with real everyday data? Earlier in the history of the field, in the first two or three decades after Benford's publication, certain researchers and theoreticians claimed that the answer lies in the purported fact that a big portion of our data actually arises from repeated arithmetic operations, most of them in the form of multiplications and divisions and, to a much lesser degree, additions and subtractions. However, it remains to be demonstrated that much of our common data truly arise from multiplications and divisions, a contention that is disputed by many. Any straightforward attempt to **explicitly** fit real-life data into such a multiplicative mold is not always obvious, and one such detailed (failed) attempt will be made in a later chapter. Claims that real-life data are **implicitly** so in some intricate non-obvious ways (which perhaps cannot be measured directly from the data itself) have been supported by some theoretical demonstrations, linking the approach to the Multiplicative Central Limit Theorem in mathematical statistics.

It is noted that geometric series, exponential growth, and exponential decay are nothing but repeated multiplication processes, albeit ones with a single (constant) factor applied. Admittedly, such logarithmic behavior appears surprising. The series keeps pacing forward, with no end in sight, leaving a long trail of logarithmically well-behaved numbers from the very start and exhibits such behavior approximately for almost any interval cut from the whole series at any two points (assuming they are sufficiently apart and/or having an integral exponent difference in the approximate). Somehow, as if by magic, the series jumps forward using carefully selected choreographic steps, not only favoring low-digit-led numbers overall, but it also does it in a way so as to yield the logarithmic almost exactly!

To take the mystery out of this seemingly bizarre digital behavior, consider 8% annual (exponential) growth with compounding from a base of 1,000. The table in Fig. 1.31 shows amounts by year and their first digits for this series. This series spends about 10 years before arriving at 2,000, but then progressively speeds up to the 10,000 mark with the overall result that it spends more time visiting low-digit-led numbers. If we focus on the addition aspect of these constant

Year	Quantity	1st digit
0	1000	1
1	1080	1
2	1166	1
3	1260	1
4	1360	1
5	1469	1
6	1587	1
7	1714	1
8	1851	1
9	1999	1
10	2159	2
11	2332	2
12	2518	2
13	2720	2
14	2937	2
15	3172	3
16	3426	3
17	3700	3
18	3996	3
19	4316	4
20	4661	4
21	5034	5
22	5437	5
23	5871	5
24	6341	6
25	6848	6
26	7396	7
27	7988	7
28	8627	8
29	9317	9
30	10063	1

Figure 1.31 Yearly 8% Growth from 1,000

multiplications, we note that progressively higher and higher quantities are being added of course. A factor of 1.08 implies adding  $\sim 80$  near 1,000, adding  $\sim 400$  near 5,000, and adding  $\sim 720$  near 9,000. It is also clear that the series repeats the same step-like digital pattern over and over again indefinitely. This explains why the series spends most of its time at low-digit-led numbers. The chart in Fig. 1.32 shows the yearly digital steps taken by the series.

Much more striking examples are given with those series where growth is so high that it exceeds the 50% or 70% range (much like the earlier sequence example of  $8 \cdot 3^N$ ). All such rapid growth result in series that seem as if they are deliberately selecting low-digit-led numbers much more often than high-digit-led

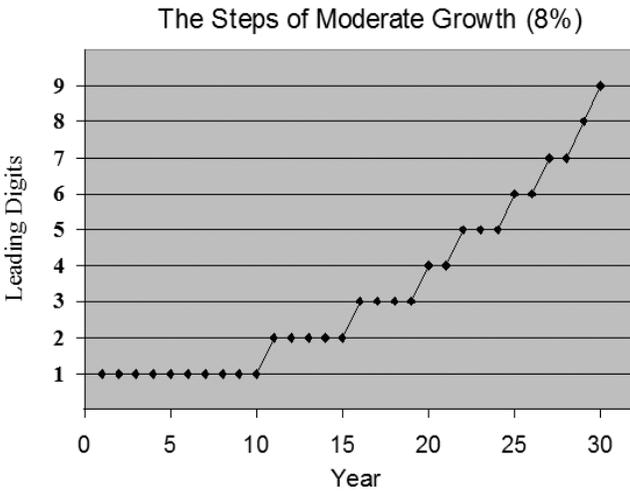


Figure 1.32 Digital Steps of Moderate 8%Yearly Growth from a Base of 1,000

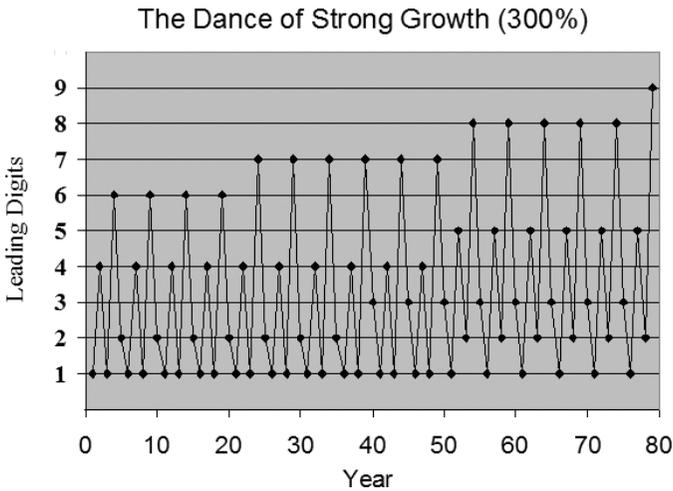


Figure 1.33 Digital Dance of the Strong 300%Yearly Growth from a Base of 1,000

ones, and doing it in a somewhat haphazard or mysterious fashion, yet consistently following the logarithmic distribution very closely! The chart in Fig. 1.33 shows the yearly digital dance taken by 300% exponential growth series starting from a base of 1,000, namely the series  $1000 \cdot 4^{\text{year}}$ .

Digital behavior of high-growth series is not without an explanation. Exponential growth is not free of units or scale but rather depends on the length

of the time used in defining the period. Banks typically quote rates per month, quarter, or year for interest-bearing accounts. Bacterial growth in the lab is typically measured in hours. Any high growth that is quoted using a particular period of time could in principle be quoted as lower growth if defined over a shorter period of time, assuming that growth is continuous, not incremental. For example, 150% annual growth is approximately equivalent to 8% monthly growth with compounding. Therefore, the supposed peculiar digital dance of rapid growth is simply the result of taking the digital pulse of numbers at points in time separated by much longer period than the one used to define equivalent lower growth. Multiple snapshots of such giant steps (taken at equally spaced time intervals) show that the series is more likely to trample over low-digit-led numbers because it spends more time treading such numbers in general, as was demonstrated earlier. For example, if  $B_H$  signifies bacteria count in hours such as  $B_0 = 4719$ ,  $B_1 = 5285$ ,  $B_2 = 5920$ ,  $B_3 = 6630$ ,  $B_4 = 7425$ ,  $B_5 = 8316$  and so forth, for **12% hourly growth**, then a daily reading of the colony relates to snapshots at the fixed hourly intervals  $B_0$   $B_{24}$   $B_{48}$   $B_{72}$   $B_{96}$ , equivalent to **1418% daily growth**. Interestingly, a truly random play on the index  $H$ , with output pointing to, say,  $B_{33}$   $B_{48}$   $B_{369}$   $B_{7538}$   $B_{9856}$ , and so forth, is logarithmic just the same, but it is no longer considered a deterministic process, rather it is considered as a random pick from a logarithmic data set. In other words, instead of daily readings every 24 hours steadily such as  $B_0$ ,  $B_{0+24}$ ,  $B_{0+24+24}$ , and so forth; varying, disorderly, and random choices of hourly intervals are selected for the times to take those snapshots of bacterial count.

It is worth noting the very general nature of multiplication processes with regards to logarithmic behavior. The factor does not have to be constant as in all the examples above. A multiplication process that starts with any random number and is derived by being successively multiplied by other random numbers shows a near perfect agreement with the logarithmic distribution, given sufficient number of sequences. It does not matter what type of random numbers are involved, be it the Normal, the Uniform, the Exponential distributions, and so forth, as long as successive multiplications are being applied, and as long as all intermediate products in the sequence are being kept and considered. We are not considering that single number of the final product containing within itself all the previous numbers as factors, but rather the initial element, the second element, the third element, and so forth. In other words, the initial element, the final element, and all the intermediate elements in between, as one large data set. An equivalent vista

for such random–multiplicative process is to think of a quantity that grows by way of a random growth rate, such as, say, 435% rapidly in the first year, 2% slowly in the second year, 17% moderately in the third year, and so forth.

Let us use formal notation for such **random-multiplicative process** for clarity.

Let  $X_i$  be a realization from any type of random variable, then the set  $\{X_1, X_1 * X_2, X_1 * X_2 * X_3, X_1 * X_2 * X_3 * X_4, \dots\}$  is logarithmic in the limit as the number of sequences grow. Here all intermediate products are retained and included in the final set to be considered. It is noted that such repeated multiplication though either dies out quickly or explodes upward very fast, namely it leads to a quick convergence to either zero or infinity depending on whether the random variable tends to be above or below the crucial value of 1. It is very important that the random nature of the above multiplicative process be acknowledged. This is not a deterministic process, even though the term ‘multiplication’ is used, deceiving many a scholar! It is also noted that this process corresponds to **Random Walk** (albeit in a multiplicative, not additive fashion), a well-known and much discussed problem in statistics and thermodynamics.

Richard Hamming, a mathematician who worked in 1945 on the Manhattan Project as a pioneer computer programmer, was assigned to calculate and investigate whether an atomic bomb detonation would ignite the whole atmosphere causing planet-wide destruction. He correctly concluded that it would not cause such ignition, and the project proceeded; its moral value and ramifications for mankind are still debatable to this day. In his seminal 1970 paper Hamming wrote about his persistent observations that repeated applications of the four standard arithmetic operations on the computer rapidly drove almost all numerical data sets towards logarithmic digital behavior, especially for multiplications and divisions. He backed up his empirical findings with rigorous mathematical reasoning and eloquent discourse. In the abstract of his paper he writes: “The paper also gives a number of applications to hardware, software, and general computing which show that this distribution is not merely an amusing curiosity.” He also writes: “The distribution is of practical as well as theoretical interest, and it is hoped that by adopting the machine’s point of view with respect to how numbers are transformed by arithmetical operations, computer scientists will become more aware of the importance of this distribution in many situations including numerical analysis.” In his paper Hamming invoked the Central Limit Theorem (CLT), an immensely important result in mathematical statistics and which also plays a prominent role

in Benford's Law and the whole digital phenomena. A direct consequence of the CLT is the Multiplicative Central Limit Theorem (MCLT) which states that a product of independent and identically distributed random variables is Lognormal in the limit. This fact in turn usually implies logarithmic digital behavior for such a product since the Lognormal distribution is nearly logarithmic whenever the shape parameter is large enough.

An obvious and important exception is when the (constant) factor itself is an IPOT number such as 10, 100, 1000, and so forth. In such cases multiplication processes are digit-neutral, and digit configuration is frozen in spite of repeated multiplications. For example, tenfold hourly growth (900%) of bacterial colony in the laboratory, with an initial count of 753 cells, implies the hourly reading of the colony count as in: {753, 7530, 75300, 753000, etc.}; digit 7 takes 100% leadership and no convergence to the logarithmic is seen whatsoever. In symbols, the series is  $B_H = 10 * B_{H-1}$  or equivalently  $B_H = 753 * 10^N$ ,  $N = \{0, 1, 2, 3, \text{etc.}\}$ ; growing with the constant factor of 10.

Very few real-life data sets come under the protective umbrella of deterministic multiplication processes such as in exponential growth series. A rare example might be found in end-of-year account balance readings of a frozen bank account untouched for, say, 50 years, without any deposits or withdrawals to disturb digital configuration. Another more realistic example might be a whole log of hourly bacteria readings in the lab for a few consecutive days or weeks. Most empirical findings of Benford's Law arise from other processes and causes, that is, they spring from other explanations (including random multiplicative processes, which play a prominent role in BL). Deterministic multiplication processes do not constitute an explanation of the whole phenomena; it only covers a very narrow class of logarithmic data sets and adds another dimension to the phenomena. Surely, other causes and sources of the logarithmic may interact and overlap each other, almost always reinforcing each other, never counteracting or neutralizing one other!

## CASE STUDY III: EXPONENTIAL 0.5% GROWTH SERIES FOR 3,233 PERIODS

---

As a case study in the logarithmic behavior of multiplication processes, an exponential slow growth series shall be digitally analyzed. The quantity growing may represent a bank account which grows with a steady stream of interest payments, or it may represent bacteria reading in a laboratory, and so forth. Initial quantity is set at 600, having a low growth rate of 0.5% (one half of one percent) per period. In order to comply with the two necessary conditions guaranteeing logarithmic convergence mentioned in the previous chapter, exactly 3,233 sequences shall be considered, implying that the last element in the series is 6009839134.9. Such length for this exponential series guarantees two results: (I) the series is sufficiently long for convergence, (II) the exponent difference between the first and the last elements is approximately of an integral value. The exponents of the first and the last elements are 2.778 and 9.779, derived from  $\text{LOG}(600)$  and  $\text{LOG}(6009839134.9)$ , respectively. Therefore the difference between the exponents is  $(9.779 - 2.778) = 7.001$ , which is very close to the integer 7.

Alternatively, the expression **Last = Constant\*First** here is calculated as  $(6009839134.9) = (10016398.6)*(600)$ , implying that the Constant term is nearly an integral power of ten value ( $\approx 10^7$ ).

To get a better sense of what data set is being considered here, the following limited snapshot from the 3233-long sequence shows the first six and the last four elements:

$$\{600.0, 603.0, 606.0, 609.0, 612.1, 615.2, \dots$$

$$\dots 5920585567.4, 5950188495.2, 5979939437.7, 6009839134.9\}$$

The 1st order digit distribution is:

Exp. 0.5% Growth — {30.10, 17.60, 12.50, 9.68, 7.92, 6.68, 5.85, 5.07, 4.61}  
 BL 1st Digits — {30.10, 17.61, 12.49, 9.69, 7.92, 6.69, 5.80, 5.12, 4.58}

The 2nd order digit distribution is:

Exp. 0.5% Growth — {11.9, 11.3, 11.0, 10.4, 10.1, 9.7, 9.4, 8.9, 8.5}

BL 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

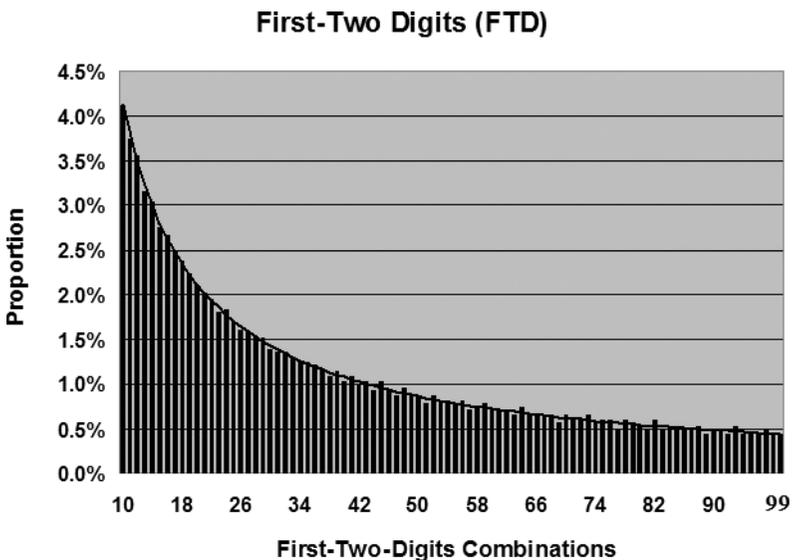
The 3rd order digit distribution is:

Exp. 0.5% Growth — {10.08, 9.93, 9.81, 10.70, 10.05, 9.46, 10.08, 10.21, 9.84, 9.84}

BL 3rd Digits — {10.18, 10.14, 10.10, 10.06, 10.02, 9.98, 9.94, 9.90, 9.86, 9.83}

1st order digits are so close to the logarithmic that they are shown here with two decimal places for better comparisons. It is extremely rare to find random data with such strong compliance with the law, perhaps it has never been observed as yet. Future data on thousands or millions of, say, exoplanets, or some incredibly large aggregations of data sets into one powerful computer database may yield comparable strong logarithmic results.

Figure 1.34 depicts the first-two-digits distribution of the exponential growth data set. Compliance with Benford's Law is almost perfect. The histogram is almost identical to the smooth continuous line of the logarithmic FTD proportion of  $\text{LOG}(1 + 1/pq)$ , therefore it almost masks it completely except for a few tiny



**Figure 1.34** First-Two Digits for 0.5% Exponential Growth Series Data Set

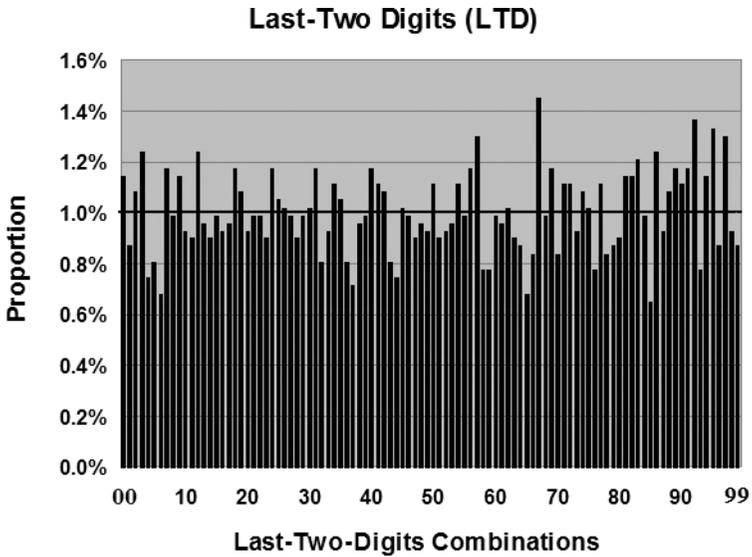


Figure 1.35 Last-Two Digits for 0.5% Exponential Growth Series Data Set

spikes. Figure 1.35 depicts the last-two digits distribution, where almost all LTD combinations come out roughly between 0.8% and 1.2%. The worst spike of 1.45% occurs at 67, but this constitutes a very mild and acceptable deviation from the theoretical 1% probability.

## CASE STUDY IV: 140 CUMULATIVE DICE MULTIPLICATIONS

---

In contrast to the deterministic nature of the previous case study regarding exponential growth series, a random multiplicative process employing the throws of dice shall be examined. This case study pertains to a newly devised casino game designed to attract well-educated and more sophisticated gamblers. The game involves the sequential throws of 140 dice by the casino. This stream of dice values is utilized in generating the series of the cumulative products of all the thrown dice. Money is bet on the value of the last 140th cumulative dice product, with the arbitrarily imposed casino rule that clients win the bet only when it is more than  $10^{70}$ . Each die has  $\{1, 2, 3, 4, 5, 6\}$  possibilities, hence in theory the lowest possible 140th cumulative product is  $1^{140}$  or simply 1, while the highest possible such cumulative product is  $6^{140}$  or  $\approx 10^{109}$ . Surely the occurrences of these two extreme cases are very rare. The casino owner has cleverly and dishonestly designed the game knowing that the 140th cumulative dice product is rarely over  $10^{70}$ , and that therefore most gamblers willing to play the game would lose.

This case study shall focus on the digital configuration of the entire set of 140 cumulative products (as opposed to focusing on the actual dice values themselves, or on the 140th cumulative dice product only).

The actual values of 140 dice faces during one particular night at the casino, in the order of occurrences are (commas omitted):

{42541225336543344365434352343136653222154636516555311522162112233416255434643523546614243135432646113253251424464224134614162331515531134645}

A very limited snapshot of the resultant 140 cumulative products, showing only the first 16 and the last 3 elements, is as follows:

{4, 8, 40, 160, 160, 320, 640, 3200, 9600, 28800, 172800, 864000, 3456000, 10368000, 31104000, 124416000, ..., 1.04635\* $10^{65}$ , 4.1854\* $10^{65}$ , 2.0927\* $10^{66}$ }

Fortunately for the casino owner, the 140th cumulative dice product came out well below  $10^{70}$ .

Digits distribution for this series of 140 cumulative products is close to being logarithmic, and this is so in spite of the very small data size of merely 140 elements. Its first- and second-order distributions are:

The 1st order digit distribution is:

Cumul. Dice product — {29.3, 15.7, 17.9, 8.6, 4.3, 5.0, 7.1, 7.9, 4.3}

Benford's Law 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

Cumul. Dice product — {10.9, 8.0, 9.4, 8.0, 13.8, 13.0, 10.9, 10.1, 7.2, 8.7}

Benford's Law 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

## THE UNIVERSALITY OF BENFORD'S LAW — TRUE IN ANY SCALE SYSTEM

---

---

What happens to digital configuration when data on human weight in kilograms, such as 55, 73, 90, are measured in British pounds and recorded as 121.25, 160.94, and 198.42? Surely any such transformation of numbers via re-scaling drastically re-arranges digital distributions. In the above example about human weight, digit 1 strongly dominates first order under the British pound, but loses heavily under the metric system.

Remarkably, this is not the case for large logarithmic data sets! If stock prices on the New York Stock Exchange quoted in U.S. Dollar are almost logarithmic, then they are also nearly logarithmic when quoted in Euros or Mexican Pesos. If time between successive earthquakes in seconds is logarithmic, then the same data measured in minutes or hours is also logarithmic! Human weight data is neither logarithmic under the British pound system, nor under the metric system, and therefore digit configuration strongly depends on the scale in use. Logarithmic data sets universally obey Benford's Law under any scale system whatsoever! The fact that the logarithmic distribution is independent of whatever scale system is used lends Benford's Law an aura of invincibility and universality! This unique property of the logarithmic distribution is called "**The Scale Invariance Principle**".

U.S. Population Centers Census is count data, requiring no scale whatsoever. Its unique non-optional 'scale' is a human being, a person living in a certain town or city. Nonetheless, the principle in a general sense implies that any multiplicative transformation by a constant value does not alter digital configuration in a significant way for any data set that is nearly logarithmic and sufficiently large. A scale change is actually nothing but a multiplicative transformation by the same fixed factor of each number in the data set under consideration. The table in Fig. 1.36 demonstrates the scale invariance principles in the case of U.S. population centers data. It shows an almost constant first-digit distribution under any multiplicative transformation. The transformations of the second and third rows by 10 and 100,

Data Set	1	2	3	4	5	6	7	8	9
U.S. Population Centers	29.4	18.1	12.0	9.5	8.0	7.0	6.0	5.3	4.6
10 * U.S. Population Centers	29.4	18.1	12.0	9.5	8.0	7.0	6.0	5.3	4.6
100 * U.S. Population Centers	29.4	18.1	12.0	9.5	8.0	7.0	6.0	5.3	4.6
4 * U.S. Population Centers	29.5	18.1	12.9	9.7	7.7	6.4	5.6	5.4	4.7
7 * U.S. Population Centers	30.2	16.9	12.6	9.9	8.3	6.7	5.9	5.0	4.4
11 * U.S. Population Centers	29.6	18.0	12.5	9.3	7.7	7.1	5.9	5.3	4.7
63 * U.S. Population Centers	30.2	17.1	12.8	9.9	8.2	6.7	5.7	5.0	4.4
0.403 * U.S. Population Centers	29.5	18.0	12.9	9.8	7.6	6.4	5.6	5.4	4.7
1.519 * U.S. Population Centers	30.4	17.1	13.0	9.7	7.4	6.6	5.6	5.3	4.8
3.693 * U.S. Population Centers	29.3	18.4	12.7	9.5	7.6	6.6	5.9	5.3	4.6

Figure 1.36 Digits of U.S. Population Centers Data are Unaffected by Multiplications

Data Set	1	2	3	4	5	6	7	8	9
Earthquake (Second)	29.9	18.8	13.5	9.3	7.5	6.2	5.8	4.8	4.2
Earthquake (Minute)	28.6	17.7	12.8	10.5	8.6	6.8	5.8	5.0	4.2
Earthquake (Hour)	29.2	16.9	12.3	9.8	8.1	7.0	6.2	5.7	4.9
Earthquake (Day)	29.4	18.2	13.6	10.2	7.6	6.3	5.4	4.9	4.5
Earthquake (Month)	31.4	16.5	11.8	9.5	8.0	6.8	5.9	5.3	4.8

Figure 1.37 The Logarithmic is Found in All Time Scales for 2012 Earthquake Data

respectively, are actually trivial, not requiring any grand principles, since digit configuration is obviously not affected by any IPOT mutiplicative transformation.

For a specific case of real-life data that does depend on scale, geological measurements data examined earlier on time between successive earthquakes during the entire year of 2012 shall be considered. The table in Fig. 1.37 gives the first-digit distributions of five different data sets of the same physical process, measured in seconds, minutes, hours, days, and months. Clearly, the variation in digital configuration is quite mild, and digital configurations are all centered tightly around the logarithmic no matter what units are used for the time dimension! Theoretically, even those small digital variations are due to the fact that earthquake data itself is not perfectly logarithmic. For the hypothetically perfect logarithmic data set (if it exists), re-scaling and multiplicative transformations have no effect whatsoever, and digital configuration is always exactly logarithmic.

## A HIDDEN DIGITAL SIGNATURE WITHIN BENFORD'S DIGITAL SIGNATURE

---

---

We conclude the first section of the book with a more detailed description of how digits obey the law. Benford's Law relates to the overall digit distribution of the entire data set in question. A closer digital scrutiny within different sub-regions reveals a very consistent pattern of an approximate **digital equality** on the left part of the x-axis for low values; the **logarithmic** property roughly around the center for the bulk of the data; then **severe digital inequality** in favor of low digits on the extreme far right for high values — skewer and even more uneven than the logarithm configuration itself. Overall, when digital configurations of all the regions on the left, right, and center are aggregated, the logarithmic is encountered as dictated by the law.

This differentiation in digital configuration along the entire range is appropriately called '**digital development pattern**', and it is found in **all** random data sets (logarithmic or otherwise) without a single exception, making it by far more prevalent and universal than even Benford's Law itself. Yet, this pattern can only be seen under a partition of the entire range along IPOT points, such as (0.1, 1), (1, 10), (10, 100), and so forth. For deterministic data types such as exponential growth series this pattern does not exist; and instead the logarithmic property is found consistently and equally everywhere throughout the entire range.

Close examination of mini digital configurations between IPOT points for the 2012 earthquake data set (measured in seconds) clearly reveals a definite developmental pattern. The table in Fig. 1.38 shows five different digital configurations. Except for the first two sub-regions on the left, skewness increases as focus shifts to the right. The general pattern seen in Fig. 1.38 is found in all other random data sets.

<b>Left Border Point:</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>
<b>Right Border Point:</b>	<b>10</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>	<b>100,000</b>
	=====	=====	=====	=====	=====
<b>Digit 1</b>	<b>8.6</b>	<b>11.3</b>	<b>15.7</b>	<b>44.0</b>	<b>98.6</b>
<b>Digit 2</b>	<b>12.5</b>	<b>10.2</b>	<b>14.7</b>	<b>23.5</b>	<b>1.4</b>
<b>Digit 3</b>	<b>18.8</b>	<b>9.8</b>	<b>13.4</b>	<b>14.1</b>	<b>0.0</b>
<b>Digit 4</b>	<b>8.6</b>	<b>10.2</b>	<b>11.4</b>	<b>7.5</b>	<b>0.0</b>
<b>Digit 5</b>	<b>13.3</b>	<b>11.0</b>	<b>10.1</b>	<b>4.9</b>	<b>0.0</b>
<b>Digit 6</b>	<b>10.2</b>	<b>12.6</b>	<b>9.6</b>	<b>2.5</b>	<b>0.0</b>
<b>Digit 7</b>	<b>9.4</b>	<b>12.1</b>	<b>9.5</b>	<b>1.8</b>	<b>0.0</b>
<b>Digit 8</b>	<b>7.0</b>	<b>10.2</b>	<b>8.5</b>	<b>1.0</b>	<b>0.0</b>
<b>Digit 9</b>	<b>11.7</b>	<b>12.7</b>	<b>7.1</b>	<b>0.6</b>	<b>0.0</b>
	-----	-----	-----	-----	-----
<b># of Data points:</b>	<b>128</b>	<b>1250</b>	<b>8234</b>	<b>9741</b>	<b>72</b>
<b>% Overall Data</b>	<b>0.7%</b>	<b>6.4%</b>	<b>42.3%</b>	<b>50.1%</b>	<b>0.4%</b>

**Figure 1.38** Digital Development Pattern for 2012 Worldwide Earthquake Data

**This page intentionally left blank**

## **Section 2**

### **FORENSIC DIGITAL ANALYSIS & FRAUD DETECTION**

**This page intentionally left blank**

## HISTORICAL BACKGROUND OF THE FIRST APPLICATIONS OF BENFORD'S LAW

---

In 1972 **Hal Varian**, a graduate student at the University of California, Berkeley, conceived the idea of utilizing Benford's Law to detect erroneous economic forecasting data, thus suggesting the first ever actual application of Benford's Law. While working at the Center for Real Estate and Urban Economics at UC Berkeley on computerized simulations of a regional forecasting project, he discovered a major flaw in the written code of the program that invalidated part of the numerical forecasts. The occurrence prompted him to think about error detection for such complex algorithms in general, and to wonder if there was a way to test whether numerical output of any given model appears 'natural' and 'reasonable' in some general statistical sense. An article he had read in *Scientific American* a few years before on Benford's Law prompted his insight to demand that computer output of future economic forecast data should obey the law (at least approximately), just as actual current data does; that 'natural' in this context is simply being digitally logarithmic! In other words, if typically (past and present) real economic data of the type of the computer simulations is approximately logarithmic, then so should be the output regarding forecasted future economic data, and that deviation from the digital law in forecasted output casts strong suspicion on the validity of the programming setup itself, indicating some human error (or possibly some intentional alteration) made in the construction of the algorithm. It is interesting to note that at that early period in the evolution of the field of Benford's Law, Varian felt obliged to write cautiously and defensively about a digital phenomenon that was not yet well-understood and lacked rigorous mathematical explanation. Thus he writes **"After all, Benford's Law is just a curious empirical, almost numerical, phenomenon"**. It is fortunate that he dared going ahead with the observable and the empirical, rather than strictly with the 'rational' and the 'explainable'.

**Charles Carslaw** from the University of Canterbury in New Zealand is the person credited with publishing the first ever paper applying Benford's Law

specifically to fraud detection in the context of financial and accounting data. In his groundbreaking 1988 paper he methodically compares the digital language used by accountants in the reporting of earnings of New Zealand firms, to the theoretical Benford proportions. His research utilizes the second order of the law to demonstrate a strong bias towards reporting earnings in excess of the psychologically important reference points of INTEGER\*I POT such as 500,000, and especially I POT such as 1,000,000. Surely when the dishonest accountant intentionally rounds up \$985,723 profit to, say, \$1,005,000, the net result is an excess of the 0 digit proportion within the second-order distribution. The reverse happens whenever the accountant is attempting to round down losses, resulting in a deficiency of the 0 digit within the second order. Carslaw then went about empirically verifying these digital tendencies in actual NZ accounting data, confirming his hypothesis. More generally, the possibility of utilizing digital comparisons exists also in other financial, accounting, and economic-related contexts, and it stems from the obvious assumption that a dishonest economist or accountant inventing numbers, such as GDP, inflation rate, interest rate, or non-existent expense and revenue data in order to manipulate investment decisions, firm's prestige, or taxes, would write those fake values without any regards to digit proportions, paying no attention to resultant digital configuration, thus ending up roughly with digital equality. Perhaps the scheming economist or the dishonest accountant may consciously assume that digits are equally distributed, as almost all people would naturally intuit, and would concoct numbers as such. The forensic method then is simply to compare digital distributions of actual accounting or economic data to that of the theoretical Benford proportions, and flag any data as suspicious if deviation is deemed too high. This forensic digital test is especially useful when the economic or financial data in question is quite large and the generic data type is known to strictly follow Benford's Law. Additional contributions to the field subsequently came with the publications by the accountant Mark Nigrini beginning in 1992, followed by C.W. Christian, S. Gupta, and more recently by Cindy Durtschi, Cleary Richard, Thibodeau Jay, and others. Such useful forensic applications have helped propel Benford's Law from a mere mathematical curiosity to an important and useful phenomenon, so much so that nowadays most governmental tax revenue departments worldwide apply it in forensic data analysis for fraud detection as their standard test. Finally, the field of Benford's Law earned its due respect when Theodore Hill in his celebrated 1995 paper gave rigorous mathematical proof validating Benford's Law in the special case of a large collection of distributions.

## METHODS IN FINANCIAL AND ACCOUNTING FRAUD DETECTION

---

---

Following the innovation of these new applications, forensic digital analysis has seen an explosion in use by accounting and auditing firms, as well as governmental tax authorities worldwide, as routine check on data for fraud. The logarithmic distribution is so ubiquitous that it is hard to overestimate its importance and relevance in forensic data analysis, as it is found in almost all financial and accounting data types. It is important to recognize the fact that each well-defined data type or distribution (logarithmic or otherwise) has its own particular leading-digit signature, a sort of hidden digital code that is not immediately obvious during the first visual preliminary inspection of the data, when the focus is on numbers and quantities rather than their digital expressions.

Maliciously invented data obviously would not obey Benford's Law, but instead typically the digits appear approximately all equally likely (uniformly distributed). At times though, the number-inventor would often repeat certain digital patterns — unknowingly perhaps in a subconscious way — reflecting a very personal preference or tendency. A cautionary flag is raised if deviation of actual from theoretical is significantly large, which calls for further scrutiny and examination of data.

First-order digits examination alone is not always the most efficient or sufficient way to detect fraud. A large spike (unusually large value) in the proportion of the first digit, say digit 8, would force the auditor to look into possibly tens or hundreds of thousands of entries, namely all amounts beginning with digit 8, and this may be too costly and time consuming. Narrowing down suspicious entries is accomplished by examining the first-two-digits (FTD) combinations, as well as the last-two-digits (LTD) combinations. For example, if a large FTD spike occurs at 85, it entails looking into only those entries beginning with digit 8 followed by digit 5, which is much easier and less costly than looking into all entries beginning with digit 8 regardless of what digit follows. Moreover, besides the issue of cost and time, FTD test also helps the auditor in finding out what exactly caused the spike, and to specifically identify the possible source of fraudulent activity. For

extremely large data sets, an examination of the first-three-digits combination is recommended in order to narrow down spikes even further.

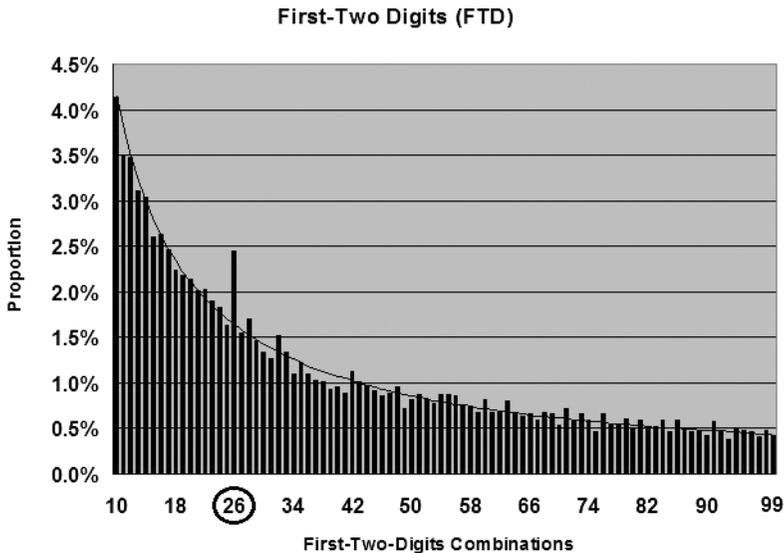
The fact that a given data set under examination conforms closely to the first-order Benford distribution does not guarantee at all compliance with the law, since higher-order digits may still be off, and numerous amounts within the data set might have been invented. Thus, adding second-order-digits test in forensic analysis is an essential part of a thorough forensic examination. This should be done as a separate test apart from the first-digits test, as well as apart from the first-two-digits combination test. An exception is made when the data set under consideration contains too few values, in which case only first-digit test is performed. This is so due to the much more delicate differences between the digits in the second order, a fact that makes second-order deviation much harder to detect in small data sets than the more prominent and obvious differences of the first order. A suggested empirical threshold in this context is to avoid second-order and FTD tests for any data set having less than 500 records. For small data sets with, say, less than 100 records, digital forensic methods in general do not work well even at the first-order level. There exists no formal statistical theory capable of giving significant threshold points for applicable sizes, and the above suggestions are totally subjective judgment from experience in dealing with data sets and forensic digital analysis.

For an extremely large account such as one with more than 100,000 or 500,000 values for example, FTD analysis might call for the examination of too many numbers thought to be suspicious, which can be too costly and time-consuming for the company. In such cases it is better to apply the first-three-digits test utilizing the theoretical probability  $\text{LOG}_{10}(1 + 1/\text{pqr})$ , followed by the value repetition test to be discussed later. This would ensure that the focus of analysis is narrowed down to fewer numbers, lowering the expense of the examination. In general, forensic digital analysis may include the following tests:

- 1) First-digits distribution.
- 2) Second-digits distribution.
- 3) Combination of the first-two-digits distribution.
- 4) Combination of the first-three-digits distribution.
- 5) Combination of the last-two-digits distribution.
- 6) Examination of first-order digital development.
- 7) Examination of second-order digital development.
- 8) Value repetition test.
- 9) Summation test.

The chart in Fig. 2.1 is the first-two-digits test performed on a hypothetical company. The continuous thin line shown is the theoretical Benford proportion of  $\text{LOG}_{10}(1 + 1/pq)$ , which falls off gradually from 4.1% to 0.4%. There is a material issue for the auditors to examine further in this accounting data set due to the substantial spike at 26. A spike is a digit or a digit combination having a significant or noticeable larger proportion than what is expected according to Benford's Law. For the particular company in Fig. 2.1, it means that there were excess amounts starting with 26, such as \$26,800, \$267, or \$260,558, over and above the logarithmic expectation. The account needs to be examined in more detail by having the auditor review all invoices relating to such amounts if that is a feasible option. The issue with the 26 spike would not have been detected easily merely with the first-digit test, which would have indicated that digit 2 occurs just slightly more frequently than it should as a first digit. A second-order test alone also would not enable us to easily detect the 26 spike. This is an example of why it is always recommended to look also into FTD, as it leads us to fraud detection more easily and also helps in directing us to the exact numbers that have caused the digital spikes.

It is the **spikes** that we seek to further investigate, not the **troughs** (unusually small values.) The reason for this is that a spike indicates that there was possibly



**Figure 2.1** First-Two-Digits Test Showing a Suspicious Spike at 26

active invention of fictitious amounts resulting in a spike, while a trough is simply a decrease in the frequency of a certain digit combination and thus does not come under any suspicion. Put another way, a number concocting scheme never directly causes a trough, only a spike if any. Admittedly, a huge spike causes all other digit combinations to come out a bit lower, but the burden of reduction is shared equally by all and thus is not really noticeable except the spike itself.

One must always tolerate of course some degree of deviation from the ideal Benford line on the FTD and LTD charts, because real-life accounting data is never exactly Benford. Yet, deviations should appear at least random and unstructured. An important measure of conformity to the law, or at least the absence of any suspicion regarding possible fraud, is having those unavoidable mini-spikes randomly dispersed from 10 to 99 on the FTD chart, and from 00 to 99 on the LTD chart, without any clustering, trend, or any repeated and systematic (cyclical) occurrences on the tens or fives for example, and so forth. For example, if all FTD combinations from 23 to 38 are below the Benford line, followed immediately by many adjacent spikes from 39 to 52 being above the Benford line, then this raises strong suspicion and it indicates that there might be some material issue with the data set. Figure 2.2 depicts one such hypothetical case where suspicious cyclical patterns are observed.

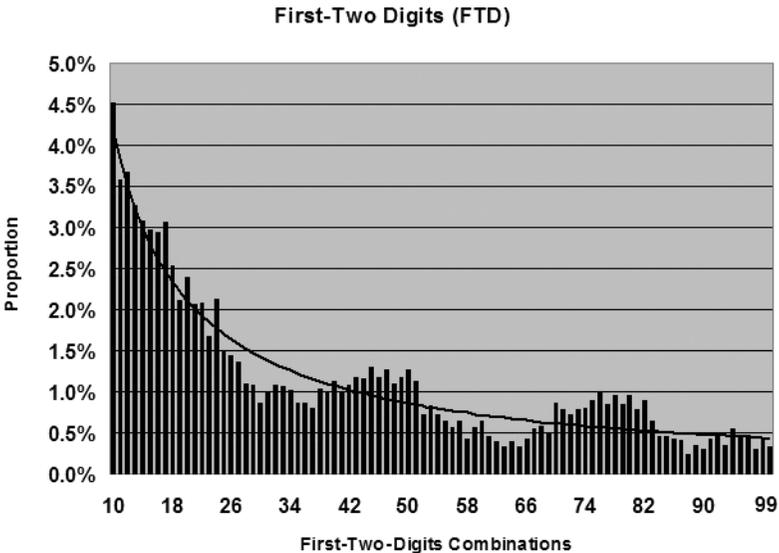
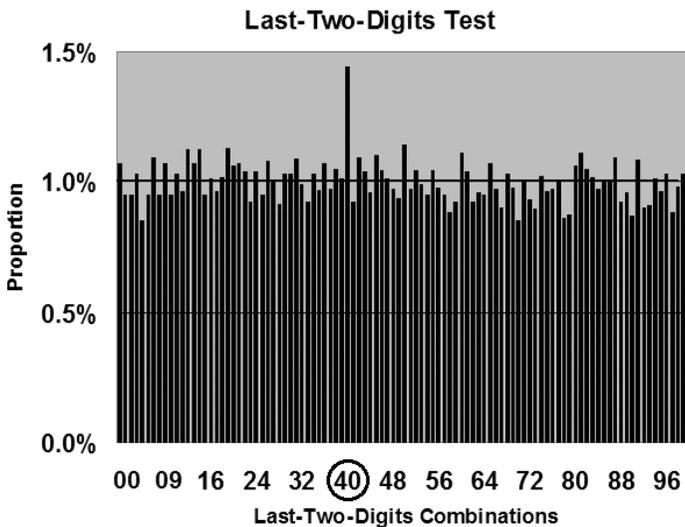


Figure 2.2 First-Two-Digits Test Showing Suspicious Structure and Cyclical Patterns

The chart in Fig. 2.3 is the last-two-digits test performed on a hypothetical company. The continuous flat line is the theoretical Benford proportion, which is steady at  $1/100$ , giving equal proportion to all 100 possibilities of  $\{00, 01, 02, \dots, 97, 98, 99\}$ . There is a material issue for the auditors to examine further in this accounting data set because of the substantial spike at 40. That means that there were excess amounts ending with 40, such as \$1,755.40, \$81.40, or \$46.40 for example. The account needs to be audited and examined in more details, especially those bills ending with 40 cents.

Last-two-digits test is a useful tool for detecting fraud because not all number inventors create values in a truly random fashion and therefore at times resultant distribution ends up not being uniform at all (as it is supposed to be). It is difficult for people to act truly randomly even when they wish or are asked to do so, and they tend to be quite creative with how numbers are being concocted. There exists a personal tendency to repeat particular numbers or number-ending combinations (cents) which are subconsciously or consciously favored. For example, a fraudster may insert numerous fake numbers combining digits 1 on the left with digit 5 on the right, such as in 1,500, 150, 10,050, and 10.50, and this can easily be detected in digital forensic testing. LTD test is an essential complementary tool in forensic analysis as it gives at times better or different indication than that seen in the FTD



**Figure 2.3** Last-Two-Digits Test Showing a Suspicious Spike at 40

test. The data analyst should keep in mind that LTD test looks for digital equality, while FTD test looks for digital disparity.

Last-two-digits test on accounting data expressing money is performed on the cents, not on the (unitary) tens of dollar amounts. On the other hand, for non-monetary integral data types (having no fractional parts) such as population count data, LTD is performed on the tens. For example, a purchasing bill in the amount of \$762.95 contributes 95 for the LTD test, while a city with a population of 675,237 inhabitants contributes 37 to LTD. A bill of \$2445.00 contributes 00 to LTD, while a city with 2445 inhabitants contributes 45 to the LTD test. The reason for choosing the cents over the tens unitary amounts is that we wish to consider digits as far to the right as possible to ensure absolute 1% equality, and thus LTD test works better on the cents rather than the dollars (the higher the order the more equal is the digit distribution). When all prices and revenues are quoted in whole dollar amounts without a single cent involved, an exception is made and LTD is performed on the tens of the monetary unit, not on the cents which are uniformly at 00. For corporate payments and revenue data in U.S. Dollars or Euros, for example, there is normally a huge spike at the 00 LTD combination. This is so because prices are often quoted in whole Dollar or Euro amounts, such as \$785.00 being price of a computer, or \$120.00 the price of a printer for example, hence a much larger portion of data ‘artificially’ ends with 00 as LTD than the theoretically predicted portion of 1%, surpassing 25% or even 50% at times (that is a whopping 25-fold or 50-fold increase over the 1% Benford proportion, and yet normal, expected, and acceptable). Particular businesses such as food chains or supermarkets also tend to quote prices ending with 99 cents, such as \$1.99 for a loaf of bread, for example, in which case LTD should show a spike at 99. At times the price list implies that there should be LTD spikes at 00 as well as at 99, and it is in such cases that LTD of accounting data should be performed on the tens as opposed to the cents. Figure 2.4 depicts one very typical LTD chart of corporate sales data. This is most typically the case in the USA, having a long tradition of pricing items just below the psychological thresholds of whole dollars, such as in \$1.99, and especially \$4.99, \$9.99, and \$99.99 which are just below the more significant barrier points of \$5, \$10, and \$100 in terms of psychologically influencing spending. In fact, frequently these two spikes at 00 and 99 are even much higher than those depicted in Fig. 2.4; it all boils down to the particular price list of the company under consideration.

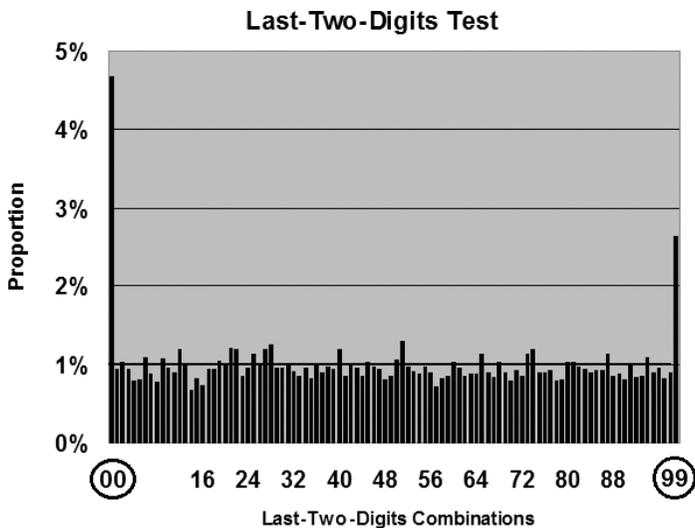


Figure 2.4 Typical Last-Two-Digits of Sales Data — Common Spikes at 00 and 99

## THE PART AND TYPE OF DATA APPLICABLE TO FORENSIC TESTING

---

Unlike typical statistical methods and practices regarding numerical data where *samples* are more economical and practical to take, or are the only possible choices in surveys and studies, in digital forensic analysis it is preferred to examine *all the data* available in the audit relating to the accounting data set in question. This is so because digital analysis is much more accurate and reliable in large data sets than in smaller ones. In order to avoid accusing honest companies of fraud, digital tests are performed only if the company has a lot of numbers. A very small company or shop with few entries would not be suitable for digital analysis by way of Benford's Law, because Type I errors would occur too frequently. A Type I error, also known as false positive, occurs when digital analysis mistakenly concludes that an honest company is fraudulent. If for reasons of high cost or secrecy, it is not possible to examine all the data, the sample should be made from all the sections of the range of the entire data set, after some thorough random mixing, without any consideration to order. Sorting the data low to high and taking portions of it as samples from specific areas leads to severe distortion in digital distributions, and not only due to the phenomenon of digital development pattern.

Digital analysis should not always include all available numbers whatsoever in an accounting audit. Often, the data analyst should eliminate from the data set prior to any digital analysis all negative values. The occasional practice to eliminate all values less than, say, 100, 50, 10, or another arbitrarily chosen low value is controversial, and it is motivated by the need to conserve time and resources, focusing solely on large amounts for which fraud is most relevant. It is usually preferable to choose an IPOT threshold value such as 10 or 100 below which all numbers are eliminated, since this usually minimizes distortion in resultant digital distributions for the rest of the data. Yet, auditors should not eliminate all values below a certain threshold if it represents a significant portion of the entire data set, say, 10% or more.

In addition, the standard procedure is never to include sums, aggregates, summaries, and totals in digital analysis of a *single company*. Rather, the analysis is

performed on the actual (original/raw) expenses, revenues, account receivables, and so forth. In other words, digital analysis for a single company must be performed on amounts at the micro level, because for the most part only such data types follow Benford's Law, not totals. Also, totals and aggregates in a sense are considered to be data duplication in the context of digital analysis and thus should be avoided.

On the other hand, when a sum, total, aggregate, ratio, count, or any statistic of a *particular accounting or financial item is pooled from a multitude of corporations* in the industry and viewed as one single data set in its own right, logarithmic behavior is definitely expected. Such an item may be almost any number, either from the annual financial statement reports, or from statistical data regarding capital markets from, say, the New York Stock Exchange, NASDAQ, and others. Examples of such items are: net income, inventory, total assets, net interest expense, trading volume, capital expenditure, total current liabilities, shares outstanding, (mutual) fund total asset, total number of employees, market capitalization, total revenues, yearly income tax expense, short interest, and others. Such a collection of numbers regarding a specific (single) issue from a large variety of companies conforms nicely to Benford's Law, given that a large number of companies are grouped together, roughly over 500–2,000 or so, depending on accuracy of compliance desired. Surely some companies do manipulate their numbers, and this causes mild deviations from the logarithmic. Yet, perhaps the combination of a financial item from so many different companies brings about some sort of offsetting effects, since those various manipulations are usually done differently or even in opposite directions, resulting in having no net effect on digit distribution in the aggregate. How could all this be applicable in forensic analysis if so many companies are combined, thus precluding the possibility of placing fault with any particular (single) company for wrongdoing? If digits are off, we do not know of course which companies are involved in fraud and which are honest, yet at least we may gain knowledge about which particular number (item) is being systematically manipulating upward or downward by a large portion of companies, (and that other financial items are left unaltered by most firms). Also, if done on a yearly basis, then a comparison can be made before and after some financial crisis, or before and after a new governmental regulatory rule regarding financial statements, accounting procedures, or capital markets. Therefore, such industry-wide digital forensic tests could confirm or disprove the claim that systematic manipulation or outright cheating takes place after (and due to) a new regulatory role.

Fusing a multiple of such item-specific data sets together as one very large super data set (encompassing many accounting and financial issues together) causes compliance with the law to be remarkably satisfactory. To mix a variety of such items from a large number of companies not only greatly enlarges the size of the data set itself, hence facilitating conformity with the law, but also further invokes Ted Hill's scheme of random pick from a variety of data sources seen in Chapter 18, thus almost guaranteeing compliance with the law.

Another possible venue is to pool together and analyze all the relevant numbers from the *financial statement of a single corporation*. Typically, results here do not closely conform to the law, because there are too few numbers in such a small data set, although there is some consistent resemblance to the logarithmic distribution, thus lending digital analysis some value. For example, if first digits of all the relevant numbers from the financial statement of a given corporation are inverted, favoring high digits, then suspicion emerges. In any case, one should not include here subtotals or other amounts that do not convey any new information, such as net income before tax, total assets, and so forth, as well as percentages or ratios (since those numbers are simply repetitions of the raw data in some sense). Yet, omitting all the above (summary/ratio) items leaves even fewer numbers for the digital analyst to crunch. Therefore only first-order test should be performed, not second-order, FTD or LTD. The same reasoning applies to yearly individual income tax statements which are considered to be even smaller data sets in the context of Benford's Law than the numbers of the financial statement of a single company. Therefore in the individual income tax case, no expectation whatsoever exists for convergence to the logarithmic and digital analysis is never performed, sparing individual taxpayers from ever being flagged for possible suspicion of fraud even if large digital deviations from Benford are detected. After a few decades or for a lifetime of tax reporting, governmental tax authorities could possibly examine digital configuration of the combined reports from all those years, for whatever it is worth.

*A large collection of monthly or yearly percent total return on investment of a single entity* such as mutual fund, real estate fund, common stock, commodity fund, stock index, risky hedge fund, and so forth, generally follows Benford's Law, but only approximately so and given that enough months, quarters or years are considered. This is so even though typically there are no more than 20, 30, or even 40 years to consider, and at most 500 monthly numbers perhaps. The data in focus is not Net Asset Value (NAV) itself, nor the factors of the changes in value, but rather the **percent** changes involved in the monthly or yearly price fluctuations, such as in

$NAV_{\text{SEPTEMBER}} = (1 + \text{percent}/100) * NAV_{\text{AUGUST}}$ . A monthly return can be thought of as the addition of 20 or so random percent values, namely the daily percent changes in the price of the fund during an entire month (since percentages are typically very low implying that compounding factor is insignificant). A typical misguided view in the literature is to liken total return to exponential growth, albeit with a fluctuating factor, but this is certainly not the case. In exponential growth series, the growing quantity (itself) is considered in its entirety, with all intermediate products retained, while total return refers only to the single resultant percent change calculated at the end of the month, quarter, or year. Total return can be considered as random walk, and as such squarely belongs to the random flavor of Benford's Law. Due to the fact that total return comes with a digital signature that is a bit similar to the logarithmic, forensic digital analysis then could serve as a powerful tool to uncover Ponzi investment schemes well before they cause havoc upon society. The recent infamous Bernard Madoff case, operating the largest Ponzi scheme in history, with an estimated loss topping 50 billion dollars, might have been stopped earlier had he been forced by law to provide more disclosure, allowing digital analysts to get clues of wrongdoing. It might be assumed that Madoff, who had been well-educated and a well-respected member of the financial elite, as well as the immediate people working with him, knew nothing about Benford's Law, thus any fake data provided by them would have probably never been concocted according to the law, and fraud could have been easily discovered with just a cursory forensic digital analysis of the fund's supposed data.

Payroll accounting data of small to medium size companies or organizations does not follow Benford's Law at all. This is so since there is hardly any meaningful spread in the data and order of magnitude is quite low (typically just slightly over one). Salaries of medium-size corporations and organizations tend to be of similar levels for the vast majority of employees, and this explains the restricted spread. If salaries for its accountants are about \$68,000, for its salespersons \$55,000, and for its computer programmers about \$72,000, while only very few managers earn more than a million and very few janitors around \$17,000, then there exists no meaningful spread in the data. It is OMV (Order of Magnitude of Variability) and its focus on the bulk of the data (central 80%) that should always constitute the criteria for compliance, not OOM (Order of Magnitude) which mistakenly encompasses everything on the margin as well as outliers leading to erroneous conclusions (see Chapter 10 for more explanations). On the other hand, when huge corporations or extremely large (governmental) organizations are

concerned, payroll data is often nearly logarithmic, since more variation exists in the pooled salary data of so many employees with varied skills and professions yielding a sufficiently large order of magnitude of variability. Nonetheless, digital development pattern is clearly seen for any payroll data, be it of large or small organizations, logarithmic or non-logarithmic, and thus a fraudulent firm or organization supplying fake payroll data can be easily detected by the lack of digital development in its provided data.

When a dishonest company wants to report strong **income** to attract investors, then amounts such as 793184 and 96.8 are rounded up and reported as say 800540 and 103.7 respectively. This happens more often when the psychological threshold of IPOT or INTEGER\*IPOT holds special significance in the marketplace. Such fake rounding would artificially increase the proportion of the 0 digit in the second and third orders. On the other hand, when a company wants to under-report **losses** it would round down amounts such as -105625 and -100.8 and report them as, say, -97885 and -98.2. This would artificially decrease the proportion of the 0 digit in the second and third orders. Hence over-representation of digit 0 indicates an attempt to increase amounts, while under-representation of it indicates an attempt to decrease amounts. In general, an excess or deficiency in the 0 digit proportion in the higher orders indicates that some fake rounding of amounts occurs. Such pioneering forensic digital test regarding fraudulent rounding performed by Carslaw was the earliest application of Benford's Law in fraud detection.

Negative and positive accounting values should be analyzed separately in any forensic digital analysis. This is so because there are different incentives to increase or decrease amounts depending on sign and motivation. A struggling company that is losing money would try to make itself look financially stronger if managers are dishonest by deflating the bottom line numbers, thus making losses appear less significant, while a successful company may try to make profits appear even stronger by inflating them to exaggerate the financial prowess of the company. In the same vein, expense data and revenue data should be analyzed separately, as the incentives to manipulate them run in totally opposite directions. Another issue which influences the direction manipulation takes place is whether the motivation is to avoid tax payments or to attract investors. Tax fraudsters would attempt to exaggerate expenses, and to under-report income. Financial-market fraudsters on the other hand would attempt to exaggerate income, and to under-report expenses in order to impress investors.

It is noted that for revenue data, relatively higher second-order proportions for digits 0 and 5 (namely spikes) are acceptable, and usually should not be a cause for concern or suspicion. This is so because often companies tend to have a lot of nice round prices such as 750, 350, 4,500, 80, 600 and so forth, which implies that random linear combinations of them yield extra proportions of digits 0 and 5 in the second order over and above the logarithmic expectation. The particular price list of the particular company investigated is an integral part of any forensic digital analysis, as it can explain or excuse large deviations from the logarithmic distribution. The same reasoning above applies to the abundance of 00 combination in the LTD chart over and above the mere 1% allocated by Benford's Law.

Another innocent cause of deviation from the logarithmic is the automatic rounding of values in the data set as an intrinsic company practice, long before any digital analysis is performed on the data. Such rounding greatly affects first- and second-digit distributions. As an example, if 0.479 is rounded to 0.5 during the process of record entry, then the frequency of digit 5 is being artificially inflated, while digits 4, 7, and 9 are being reduced in the first, second, and third orders. Obviously, such rounding innocently affects all higher-order distributions.

For a company where all prices end with 99 cents (most typically occurring in gas stations all over the USA), deviation of the last-two-digits configuration (LTD) from Benford could be quite mild in spite of the 99 bias. This is so since a variety of different multiples of 99 cents yield other digit combinations. For example, a purchase of 13 gallons of gasoline at \$3.99 per gallon would cost \$51.87, while 21 gallons would cost \$83.79, and LTD combinations do not end up as 99. An exception is found though whenever cents are exclusively 5 or 10, because then all possible multiples still yield the same closed set of 5 and 0 as the very last digit of LTD combinations. In any case special attention should be placed by the digital analyst to any possible extra charges, such as sales tax which tends to significantly alter resultant digital combinations. For example, even if all prices are quoted in 5 or 10 cents, a sales tax of 3% would cause other LTD combinations to occur.

There are many examples of **honest number inventions** that have nothing to do with fraud, such as gift amounts that corporations donate to charity organizations. For example, a \$25,000 gift from IBM to educational computer programs in schools is an honestly invented number. Particular prices invented to entice potential buyers to maximize profit, such as gasoline at \$3.99 per gallon and so forth are another example. Surely such numbers do not follow Benford's Law. Another very common example of honest number invention is investment deposit checks written by millions of individuals to banks and financial institutions, typically in

amounts of whole hundreds or whole thousands of dollars. As an example, when John Smith writes a personal check of \$3,000 when buying into some stock mutual fund, the number 3,000 has been honestly invented.

In the event in which a fraudulent company multiplies all its (original and real) revenue numbers by a constant factor such as, say, 1.25, to artificially inflate income by 25% in order to impress investors and the capital markets, this can never be detected by digital analysts. The scale invariance principle guarantees logarithmic behavior for the multiplicatively transformed fake data, and so if the original revenue data obeyed the law before the fraud took place, the newly created fraudulent data is unfortunately logarithmic just the same! It takes concocting fake numbers one-by-one in the minds of the fraudsters (without consistent reference to real-life existing data) for Benford's Law to be applicable in fraud detection.

Evidently, any fraud due to unreported dealings where no transaction is actually recorded in the company's database, such as bribes, kickbacks, asset thefts, and so forth, cannot be detected by digital analysis. In addition, few (rare) accounting data types do not obey Benford's Law to begin with and therefore cannot be so tested, such as payroll (salaries) amounts, amounts with an arbitrary minimum or maximum, amounts that are influenced by human thought like ATM withdrawals, fixed prices, and so forth.

Election results closely conform to Benford's Law. This is so since electoral results are simply manipulated fractional population data; as in 31% of population voting for candidate A promising policy bundle X, 27% voting for candidate B promising policy bundle Y, and the remaining 42% of the population simply not voting in order to preserve their human dignity and freedom by not participating in such a ridiculous charade, patiently awaiting the triumphant arrival of direct and participatory democracy instead. Since population data itself conforms strongly to the law, by the scale invariance principle, the same conformity should be found here for fractional population values. A more careful examination of the underlying statistical process reveals that this is ultimately a Random Linear Combination from random logarithmic data. This confers election results an even more logarithmic aura than mere population data. A cursory look at details of election data reveals this, as it typically reads: 43.78% Democratic vote in Rockland County, New York; 50.01% in Palm Beach County, Florida; 23.95% in Mariposa County, California; and so forth. A few fascinating electoral studies using Benford's Law to detect potential electoral fraud have been published recently, with better results obtained by way of the second-order-digit distribution.

Forensic digital analysis for the purpose of fraud detection can be extremely useful even when the data set under consideration is not Benford at all. **Each statistical process or particular real-life data type has its own unique leading digits signature, thus two different manifestations of the same process or data type can be digitally compared with the expectation that their digital configurations would show strong similarity.** Obviously, besides Benford's Law itself, there are in principle infinitely many other digit distribution laws governing the infinitely many different possibilities of statistical processes, distributions, physical systems, and data types. Benford's Law is certainly the most important one due to its incredible prevalence. Two manifestations of the same data type could arise for example from a single company with snapshots at two different instances in time, or it could arise simply in the comparisons between two different plants, offices, or branches, as well as two different companies within the same industry regarding the same data type. An example of such comparative study may be the digital analysis conducted by the general manager of a very large firm for the purpose of forensically studying payroll data for all its 76 branches around the country. Even though payroll data is not logarithmic at all, digital distributions in all its 76 branches should have quite similar configuration, assuming an approximate uniformity of operations. If the branch in New Orleans, for example, shows a distinct digital configuration for payroll as compared with all other branches, then it should be further investigated. In another example, a new regional payroll officer was appointed, say, at the beginning of 2005 to lead the branch in Chicago, and the general manager at the headquarters is interested in comparing 2004 and 2005 digital configurations of Chicago payroll data to potentially uncover corruption or outright theft by the newly appointed payroll officer.

Serial numbers, code numbers, index numbers, and such, are not logarithmic, and their digital distributions are for the most part uniform, as each of the ten digits is randomly selected with equal 10% probability. Artificial numbers invented for particular situations, such as ATM withdrawal numbers, are also not logarithmic, and their digital configuration is a function of the particular set of numbers invented. Lastly, numbers with a built-in maximum or minimum value, either artificially human-made or due to some natural barrier to values, are also not logarithmic.

## CASE STUDY V: U.S. MARKET CAPITALIZATION ON JANUARY 1, 2013

---

Any particular accounting or financial item pooled from a multitude of corporations strongly complies with Benford's Law. As a case study, Market Capitalization (in units of million U.S. Dollars) as of January 1, 2013 in a list of 5,486 U.S. companies shall be digitally analyzed. The data can be downloaded from New York University website:

[http://pages.stern.nyu.edu/~adamodar/New\\_Home\\_Page/data.html](http://pages.stern.nyu.edu/~adamodar/New_Home_Page/data.html).

The file to download is labeled as row "Current (January 2013)" and as column "U.S.". Column J contains Market Capitalization. Out of 6,177 total entries, 691 entries of zero value are eliminated, yielding 5,486 positive values. The rather large data size of 5,486 values should almost guarantee compliance with the law. It is essential to further eliminate all 307 values below \$1.00 for the second order as well as for the FTD test. This is so since all values are rounded here to one decimal place, and therefore all values less than \$1 appear as \$0.20, \$0.70, \$0.40, and so forth, rendering their second digit meaningless as it is artificially fixed as 0 digit. This elimination leaves 5,179 values applicable for the second order and the FTD tests, which is still quite large enough in a statistical sense.

The 1st order digit distribution is:

USA Market Capitalization — {30.1, 18.0, 12.6, 9.7, 8.1, 6.5, 5.9, 4.6, 4.4}

Benford's Law 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

USA Market Capitalization — {11.8, 10.6, 11.1, 10.1, 10.3, 9.3, 9.2, 10.2, 8.5, 8.8}

Benford's Law 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

Both, first as well as second digits distributions came out very nearly logarithmic. As noted by Carslaw, dishonest companies attempt to present their capital situation in a more positive way by inflating revenue and income amounts just slightly over IPOT thresholds points, such as when 936,474 is reported as 1,004,653, causing

digit 0 to gain extra proportion in the second order. Yet in this particular context one must bear in mind that companies do not have any/much control over their Market Capitalization which is defined as (Current Stock Price)\*(Shares Outstanding), hence Carlsaw's observation cannot be applied here, and this fact is confirmed by the modest proportion of 11.8% for the 0 digit in the second-order-digit distribution.

The chart in Figure 2.5 depicts the first-two-digits distribution of U.S. Market Capitalization. This FTD chart is within a very reasonable range of Benford's FTD curve shown as a continuous falling line, and as evident by the absence of any noticeable large spikes here.

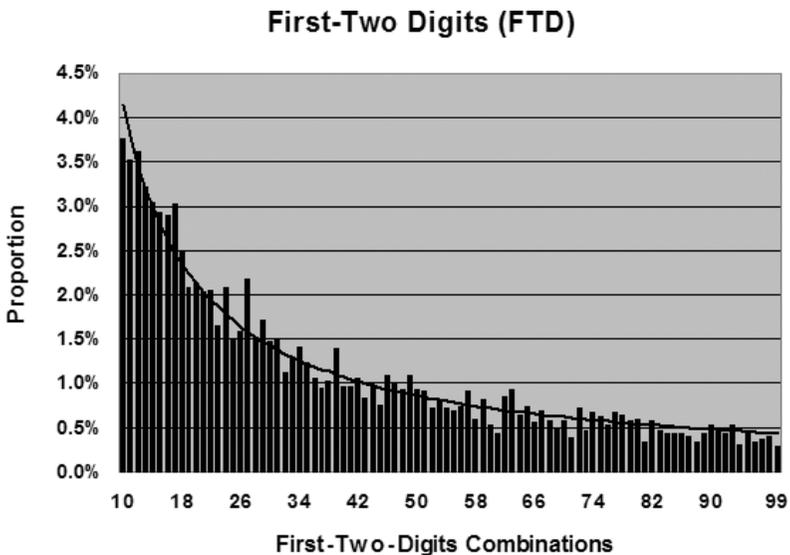


Figure 2.5 FTD Chart of Market Capitalization for 5,179 U.S. Companies

## CASE STUDY VI: MICROSOFT CORPORATION FINANCIAL STATEMENT

---

The collection of all the relevant numbers from the financial statement of a single corporation does not follow the law, yet it has some similarity to the logarithmic. As a case study, we examine the Q3 2010 Financial Statement (FS) of Microsoft Corporation. Several values such as sub-totals, ratios, etc. were excluded, leaving only 78 raw numbers for digital analysis. For such small data set only first digits should be analyzed, although we shall dare to glance for a brief moment at the second-order-digits distribution as well. First-two-digits and last-two-digits distributions should certainly not be included here due to the very small data size.

The 1st order digit distribution is:

Microsoft FS — {24.4, 11.5, 12.8, 6.4, 16.7, 7.7, 6.4, 6.4, 7.7}  
 Benford's Law 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

Microsoft FS — {11.5, 17.9, 12.8, 9.0, 9.0, 2.6, 17.9, 7.7, 7.7, 3.8}  
 Benford's Law 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

Although deviation from the first-order logarithmic is substantial, there is some strong similarity. Such resemblance to the logarithmic where low digits dominate overall distribution and high digits are allocated less leadership is very typical in FS data of a single company. Rarely does an honest corporation present its FS data anywhere near digital equality or skewness in favor of high digits. Hence this particular feature of the number-anemic FS data can be utilized to flag any corporation as possibly fraudulent if it presents digital configuration quite differently than the norm, with either digital equality or inverted skewness. Yet auditors and digital analysts must keep in mind that this should be taken with a large grain of salt due to the very large variation of this result which relies on a very small data size. Hypothetically one can make an empirical study encompassing thousands of financial reports from numerous corporations to arrive at some more specific digital guidelines. Future research may lead to this direction perhaps. Figure 2.6 depicts a

condensed version of Microsoft financial data, where DELETE signifies deleting a sum or a sub-total and such in the calculation of digit distribution. The text-less part on the right (representing other relevant financial numbers) is depicted without any descriptions due to scarcity of space. All 78 values are shown in Fig. 2.6.

## Microsoft Corporation Balance Sheets

<b>Assets</b>			
Current assets:		\$ 16,195	\$ 4,785
Cash and cash equivalents	\$ 8,161	3,139	3,959
Short-term investments	<u>36,012</u>	2,196	527
Total cash, cash equivalents	<i>DELETE</i>	<u>2,806</u>	5,126
Accounts receivable	9,646	938	1,795
Inventories	1,242	<u>58</u>	<u>3</u>
Deferred income taxes	2,344	7,116	
Other	<u>2,176</u>	<u>114</u>	\$ 3,323
Total current assets	<i>DELETE</i>	<u>1,820</u>	1,630
Property and equipment	\$ 7,771		(560)
Equity and other investments	9,211	\$ 0.63	3,388
Goodwill	12,471	\$ 0.62	382
Intangible assets, net	1,077	8,614	<u>(1,047)</u>
Other long-term assets	<u>1,429</u>	8,695	
Total assets	<i>DELETE</i>	\$ 0.16	61,935
		<u>5,505</u>	<u>(14,993)</u>
		\$ 5,410	
<b>Liabilities and stockholders' equity</b>			
Current liabilities:		694	\$ 814
Accounts payable	\$ 3,654	528	4,721
Short-term debt	1,000	(29)	(814)
Accrued compensation	2,252	(5)	177
Income taxes	2,136	(148)	(4,399)
Short-term unearned revenue	12,767	5,881	(1,118)
Securities lending payable	909	(6,862)	5
Other	<u>3,139</u>	3,674	<u>(25)</u>
Total current liabilities	<i>DELETE</i>	(468)	
Long-term debt	9,665	208	\$ (564)
Long-term unearned revenue	1,152	62	(7,417)
Deferred income taxes	540	(400)	870
Other long-term liabilities	<u>7,384</u>	(911)	1,427
Total liabilities	<i>DELETE</i>	<u>560</u>	<u>727</u>

Figure 2.6 Raw Data from Microsoft Q3 2010 Financial Statement Report

## CASE STUDY VII: TOTAL RETURN OF ATHENA GUARANTEED FUTURES FUND

---

A large collection of monthly or yearly percent total return on investment of a single entity such as mutual fund generally does not follow Benford's Law, but it is close to it in the approximate. As a case study, we shall digitally analyze monthly total returns of Athena Guaranteed Futures (AGF) Ltd Fund managed by Man Investments, from its inception in 20/12/1990 until 31/01/2013. The data set provides 266 monthly returns. By taking the absolute value of all the monthly returns, negative values are incorporated together with all the positive values as a singular data set. The data can be downloaded from Man Investment website [www.man.com](http://www.man.com).

The 1st order digit distribution is:

AGF — {19.6, 16.2, 16.6, 14.3, 9.8, 6.8, 6.4, 4.9, 5.3}  
 Benford's Law 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

AGF — { 8.7, 12.5, 10.6, 7.2, 12.8, 11.7, 10.6, 12.8, 6.8, 6.4}  
 Benford's Law 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

While not logarithmic, first-digit distribution is also not too far from it as low digits dominate overall distribution while high digits are allocated by far less leadership, and this is quite typical in most total return data sets — given sufficient number of periods. Figure 2.7 depicts a small portion of AGF Fund data. The focus is on the last column “Monthly Change” which is the one being digitally analyzed. The fund has done quite well during the past two decades growing by more than tenfold, and as far as the digital analyst can tell this is all very real, financially sound, and quite honest. Although percent is shown with only one decimal place, it has further digits to the right and so second-order digits can be found for all periods.

<b>Date</b>	<b>NAV(USD)</b>	<b>Monthly Change</b>
<b>Inception</b>	<b>10.0</b>	
<b>December 31, 1990</b>	<b>9.87</b>	<b>-1.3%</b>
<b>January 31, 1991</b>	<b>9.63</b>	<b>-2.4%</b>
<b>February 28, 1991</b>	<b>9.61</b>	<b>-0.2%</b>
<b>March 31, 1991</b>	<b>9.65</b>	<b>0.4%</b>
<b>April 30, 1991</b>	<b>9.41</b>	<b>-2.5%</b>
<b>May 31, 1991</b>	<b>9.37</b>	<b>-0.4%</b>
<b>June 30, 1991</b>	<b>9.41</b>	<b>0.4%</b>
<b>July 31, 1991</b>	<b>8.97</b>	<b>-4.7%</b>
<b>August 31, 1991</b>	<b>9.31</b>	<b>3.8%</b>
<b>September 30, 1991</b>	<b>9.75</b>	<b>4.7%</b>
<b>October 31, 1991</b>	<b>9.81</b>	<b>0.6%</b>
<b>November 30, 1991</b>	<b>9.87</b>	<b>0.6%</b>
<b>December 31, 1991</b>	<b>11.2</b>	<b>13.5%</b>
<b>January 31, 1992</b>	<b>10.22</b>	<b>-8.8%</b>
<b>February 29, 1992</b>	<b>9.96</b>	<b>-2.5%</b>
more data here for 1992–2012		
<b>August 31, 2012</b>	<b>125.09</b>	<b>-2.3%</b>
<b>September 30, 2012</b>	<b>125.15</b>	<b>0.0%</b>
<b>October 31, 2012</b>	<b>121.22</b>	<b>-3.1%</b>
<b>November 30, 2012</b>	<b>123.86</b>	<b>2.2%</b>
<b>December 31, 2012</b>	<b>123.88</b>	<b>0.0%</b>
<b>January 31, 2013</b>	<b>125.32</b>	<b>1.2%</b>

**Figure 2.7** Monthly Total Returns for Athena Guaranteed Futures Ltd

## ESTABLISHING DIRECT CONNECTION BETWEEN DIGIT ANAMOLY & FRAUD

---

Whenever possible, it is important to connect directly suspicious and abnormal digital distributions found in the data to some specific incentive to cheat. Unfortunately this is not normally the case, and the forensic analyst must dig further to find the source of the fraud. For example, if too many 12, 13, and 14 first-two-digits combinations (FTD) are found in the data, but very few 15, then perhaps there is something especially negative or costly in amounts of or exceeding say \$1,500, \$150,000, or \$15 million, such as perhaps a penalty, a fee, a surcharge, or a higher tax rate. Perhaps large amounts over say \$150,000 trigger an automatic audit or necessitate a formal report to the CEO or tax authorities, a report or an audit which the fraudster faking the data obviously is wishing to avoid and therefore numbers concocted are crowding out near the 15 FTD mark (being that the fraudster wishes to create them as large as possible but without detection). An instructive example of such direct connection between digital spikes and specific incentives to cheat is demonstrated in the case of Real Estate Tax of some very large municipality, regarding its data set on reported house prices. Being highly progressive and fair in its tax policy, the municipality's schedule of its Tax Rates is shown in the table of Fig. 2.8. The unique feature of this tax schedule is that by 'crossing' the, say, \$100,000 mark, the owner pays a higher rate not only for the part above \$100,000, but also for the basic part below it, and this gives the owner a great deal of incentive to cheat. This is so since the law enforces a 2% tax rate on the entire value of any house whose price is between \$100,000 and \$999,999.

In as much as house owners are capable of influencing the official book value of their houses, say by paying bribes to appraisal specialists, or by falsifying some documents, multiple digital spikes near the end of the first-two-digits (FTD) chart (i.e. 99) are expected to be found by the forensic analyst. Here the digital analyst can easily and directly connect digital abnormalities with the fraud committed. One hypothetical example of such a chart can be seen in Fig. 2.9 which refers to the (officially reported) data set pertaining to house values throughout the

House Value	Annual Tax Rate
Below \$99,999	1%
\$100,000 - \$999,999	2%
\$1,000,000 - \$9,999,999	5%
\$10 million +	10%

Figure 2.8 Real Estate Tax Rate in a Progressive Municipality

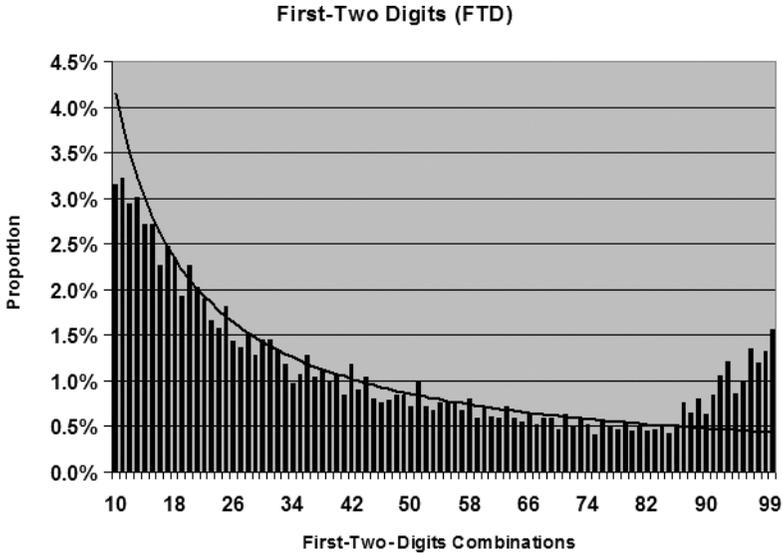


Figure 2.9 Concentrations of FTD Spikes and Troughs for the House Value Data

municipality. The basic assumption underlying digital analysis here is that data on house values in a given large municipality is Benford. A house worth \$105,000 would be falsified as being worth only \$98,000, say, while another house worth \$1,240,000 would be made to look as if it's worth only \$960,000. A huge reduction in due tax is thus achieved, as tax bracket falls sharply and dramatically, even for an incremental and small percentage reduction in price. The net effect of all this on digital proportion would obviously be the sharp reductions in 10, 11, 12, 13, etc. FTD combinations, and noticeable spikes around 89 to 99, or perhaps just around the shorter range of 95 to 99, and so forth. Note that typically a house worth \$105,000 would not be falsified as being worth only, say, \$50,000, even though that would reduce tax due even further, since such large fraud is difficult to commit and also can be quite easily detected, while changing the price from

\$105,000 to, say, \$97,000 is easier to do and it is almost unnoticeable. All this implies that if such fraudulent changes are frequent and widespread in the municipality due to lax regulations and lack of control, it would crowd out FTD chart towards the high end of 99, leaving the first few combinations near 10 with significantly lower proportions. Therefore Fig. 2.8 explains Fig. 2.9, and a direct connection between abnormal digital distributions and specific incentive to cheat has been established. Surely, the municipality would have been better off creating a more refined tax schedule having an equivalent overall tax rate, because such a schedule would reduce incentive to cheat. For example, a more gradual increase in rates in steps of \$10,000 would prevent large tax losses for the municipality, saving millions of dollars. Curiously for this Real Estate Tax case, the last-two-digits (LTD) test cannot provide any additional information about possible fraudulent house value reporting. For example, the fraudulent change of \$105,000 house value to \$98,000 leaves LTD totally unaffected at 00, and provides no clue of any wrongdoing by the owner.

Another example of such clear-cut association between digital abnormality and specific incentive to cheat can be found in the requirement on the part of the U.S. Internal Revenue Service obliging all financial institutions to send special reports relating to transactions involving cash whenever deposits or withdrawals exceed \$10,000. Excerpt from [www.irs.gov](http://www.irs.gov) reads: *“Generally, any person in a trade or business who receives more than \$10,000 in cash in a single transaction or related transactions must complete a Form 8300, Report of Cash Payments Over \$10,000 Received in a Trade or Business (PDF). Form 8300 is a joint form issued by the IRS and the Financial Crimes Enforcement Network (FinCEN) and is used by the government to track individuals that evade taxes and those who profit from criminal activities.”* Hence, first-two-digits chart (FTD) of all cash transactions in the USA, or within just a single large bank, should show similar although less intense trends of spikes and troughs seen in the above case of municipality real estate tax. Although here the particular amount of \$10,000 would be dwarfed by numerous other more innocent FTD-transitional amounts such as \$100, \$1000, and \$100,000 which have no relationship at all to the \$10,000 threshold rule (thus blurring the effect and rendering it a bit hard to detect on the FTD chart). Unlike the municipality case with three FTD-transitional points which reinforce each other, here there is only one such point. Hence, in the IRS reporting case, the forensic analyst would be better off detecting the avoidance of the \$10,000 mark by examining the actual histogram of the data itself, skipping digital analysis altogether. The suggestion given to the forensic analyst in the

municipality case to examine FTD emanates from the fact that on the FTD chart the confluence of these three transitional points reinforces those trends and this leads to much better detection by way of digital analysis than by merely examining the actual data histogram itself with its more gentle and milder deviations around the three different IPOT points.

A third example with a clear-cut connection between digital abnormality and specific incentives to cheat is given by Christian and Gupta (1993) relating to the IRS Personal Income Tax Tables within the standard Form 1040. Here, tax due increases ‘abruptly’ in steps for each additional \$50 of income, thus giving the dishonest taxpayer an incentive to fake his or her income and to reduce it slightly. An excerpt from the tax code on the form reads roughly: “*If line 43 (taxable income) is at least \$7,300 but less than \$7,350, your tax is \$733 — if you are single*”. And: “*If line 43 (taxable income) is at least \$7,350 but less than \$7,400, your tax is \$738 — if you are single*”. In this case, a lowering of the tax obligation occurs at times even via a reduction by a single dollar, which at a marginal tax rate of 10%–30% is a bit significant for the truly frugal person, namely a sure \$5–\$15 savings in tax payments. If John Smith truly earned \$7,350.50, but dishonestly reports his income as \$7,349.50, he would be cheating the tax authorities by \$5. This dishonest tendency in relatively large segment of the taxpayer population can be easily discovered by the digital analyst. As opposed to the two earlier examples of municipal tax data and large cash transactions where the focus was on FTD distribution, the focus here should be on the LTD distribution (on the whole dollar amounts excluding cents), since FTD distribution is totally irrelevant in this case. The combined (income) data from thousands or perhaps millions of taxpayers are mashed together into one single file, and last-two-digits test (LTD) is performed on the dollar amounts (omitting cents altogether), where spikes abound around 45–49 and 95–99, while troughs abound around 00–05 and 50–55. Here, the confluence of so many transitional points merged onto a single LTD chart strongly reinforces all these individual effects, and this greatly helps in observing with clarity the effect of such fraudulent tax reporting; while on the very long data histogram itself, this tendency is greatly diluted and may not even be observed at all (unless all taxpayers cheat in unison.)

## POST-TEST CONCLUSIONS

---

---

If a company's data is found **not** to obey Benford's Law then there are four possibilities:

1. False positive. The company is **honest**, but by some rare random statistical chance its data deviated from Benford.
2. The assumption that this particular generic type of accounting or financial data follows Benford's Law is not true, and the company is actually **honest**.
3. The company has some very particular business configuration that causes its data to deviate from the logarithmic. For example, in the case of revenue data, some items for sale may cause its revenues to favor certain digits, such as when its main product is a laptop priced at \$439 and first digits 4 and 8 (from the sale of two laptops) are over-represented. For expense data, for example, when a standard ton of raw material is priced at \$7,496, digit 7 would be over-represented in the first order. Here, digital tests indicate possible fraud, but the company is actually **honest**.
4. The company is **dishonest**, and its data is fraudulent — faked by its accountants.

When a company is strongly suspected of fraud due to a decisive failure to pass Benford's digital tests, we still do not know which entries were fake and which were honest. Perhaps all were fake, but all we know is that the data in its entirety is likely to be fraudulent. This is certainly a serious drawback in fraud detection via digital analysis. An even more serious pitfall occurs when a company has tens of thousands of entries in its accounting data and gave only a few fraudulent numbers, say, 23, in which case it would be impossible to discover the fraud via digital analysis. Fraud detection by way of digital analysis can only be done if a significant portion of the data was fraudulently invented. The challenge arises in such extreme cases of false negative when these 23 fake numbers are of very high values, representing a large portion of money involved. Fortunately, most fraudulent companies

and their dishonest accountants do not know about the existence of Benford's Law and so may prefer to cheat by changing thousands of numbers instead of only 23, mistakenly thinking that by spreading the fraud around many numbers it would not be detected. Actually, by spreading the fraud around so many numbers, fraudsters actively enable those zealous and righteous digital analysts to find the fraud involved!

## DETECTING FRAUD VIA DIGITAL DEVELOPMENT PATTERN

---

---

Two serious pitfalls arise in the context of forensic digital analysis and fraud detection. The first is when the data itself is not inherently Benford to begin with and thus cannot be so tested (unless it possesses its own unique digital signature as discussed earlier near the end of Chapter 27). Almost all accounting data types follow Benford's Law with the pronounced exception of payroll data and very few other cases. The second pitfall is whenever fake data is invented and provided by the sophisticated and well-educated cheater already aware of Benford's Law and all its higher-order features as well. The latter difficulty is a factor that will become increasingly more problematic in the future, and will represent a serious challenge to forensic data analysis applying Benford's Law when accountants and executives gradually become aware of this digital phenomena and will attempt to calibrate digits in fake data according to the law so as to make it appear genuine.

Knowing that random data evolves and develops its overall Benford property along the entire range in a very particular way can be applicable in forensic digital analysis for the two scenarios above. Firstly, accounting data that does not conform to Benford's Law such as payroll data and others should still show clear graduation from an approximate digital equality or even predominance of high digits on the leftmost part of the data, to the logarithmic around the center, culminating in severe inequality on the far right. This is so since they are inherently random processes, and all random data (Benford and non-Benford) appear with digital development pattern. The absence of such development pattern in forensic tests should trigger suspicion on the part of the auditors and data analysts and would merit further investigation, especially if digit distribution is shown to be very consistent and roughly uniform across IPOT sub-intervals. Secondly, whenever fake data is invented by a sophisticated cheater well-aware of Benford's Law, it can be assumed that such a clever cheater would nonetheless be unaware of the patterns of development in leading-digits distributions and thus would not be able to truly mimic real-life data with all of its intricate and hidden inner properties. Unknowingly, he

or she would naturally concoct data in such a manner where the Benford property is true, consistent, and steady throughout the entire data, mistakenly creating deterministic-like data type. Another possibility is that such a sophisticated cheater would unknowingly concoct fake data with another developmental pattern style that is neither the steady deterministic one nor the gradual random one, or that data would show a meaningless zigzag pattern-less style. The theoretical digit development pattern universally seen in random data sets is expected to be forensically found and confirmed only if data is honestly reported.

**The manner and details of how digits end up being logarithmic in the aggregate over the entire range of the data set is akin to a hidden signature within a signature, namely the hard-to-forge development signature secretly written deep within that Benford's own digital signature of  $\text{LOG}(1+1/d)$ .**

Detailed methods and numerical algorithms for the precise measurement and detection of digital development pattern in data shall be given in the next section on Data Compliance Tests.

## THE DILEMMA OF FTD VERSUS LTD FOR DIGIT-ANEMIC NUMBERS

---

---

When a number is blessed with lots of non-zero digits such as **96,378.25** for example, it obviously contributes **96** for the FTD test, and **25** for the LTD test. But what shall be done with digit-anemic numbers such as **73**? Should we incorporate the digit combination **73** in FTD test, or in LTD test? Surely the first 2 digits are **73** here, but it is also true that the last 2 digits are **73** just the same! Yet, it is not proper of course to place a digit or digits combination on both FTD and LTD charts. FTD depicts disparity and skewness, while LTD depicts equality — and as such they convey two contradictory messages.

For data with numerous three- or four-digit numbers, namely data with many ‘short’ numbers having few digits, LTD test should not be applied, because proportions here mostly follow the gently skewed results of Benford’s Law applied to the second and third orders, not the perfectly equal 1% distribution of LTD. For data exclusively with two-digit numbers, last-two-digits distribution may actually follow the classic first-two-digits skewed distribution of Benford if data is deemed to be logarithmic in spite of the fact that it has very little spread (i.e. small order of magnitude).

This is the important issue that arises in the context of FTD and LTD tests regarding numbers valued less than 10 with only a single non-zero digit, such as 0.04, 0.5, 0.0007, 8, and so forth, as well as with numbers valued less than 100 with only two non-zero digits such as 3.4, 0.0081, 59, 0.067, and so forth. More specifically, should we disregard the data value 8 in the context of FTD test altogether since the minimum FTD combination is 10 (i.e. second order is totally missing for the data value 8 — the second digit does not even exist!), or should we consider it simply as 8.0 and include it in FTD with other numbers beginning with 80? If we do include 8 in the FTD test, should it also be included in the LTD test? And if so, should it be added as 08 or as 80 LTD combination? The discussion about FTD versus LTD tests and the elaborate details of their procedures distract us momentarily from the underlying law itself which strictly dictates everything

and so we tend to lose sight. Surely the first digit of the number 8 is 8; it comes with 5.1% probability as in  $\text{LOG}(1+1/8)$ ; and it strictly belongs to the skewed FTD camp if any, as opposed to the 1% equality of the flat LTD chart! The fact that second order is missing for the number 8, or that 8 needs to be seen first as 8.0 in order to invent a second digit, may disqualify it from FTD, but it has no place whatsoever in the LTD camp! The standard procedure currently in accounting and auditing circles whenever there is a large portion of such low-valued and digit-anemic numbers in the data set is not to ignore them, but to multiply all the numbers in the data set by an IPOT factor such as 10 or 100, so as to artificially 'add' some 0 digits on the right (i.e. moving the decimal point sufficiently to the right) and then to be able to place them squarely at least with the FTD test. For example, if the lowest such digit-anemic number is 0.006, the entire data set is transformed by multiplying each value by 10,000, with 0.006 metamorphosing into 60. The author would like to criticize such an approach for generally distorting some digital configurations (especially high orders and LTD), and would like to suggest instead incorporating single-non-zero-digit number such as 8 and 0.006 only in the first-order test, eliminating them altogether before performing the second-order, FTD and LTD tests. For pairs-of-two-non-zero-digits number such as 59, and 0.092, the author would like to suggest incorporating them only in the first, second, and FTD tests, and eliminating them before performing LTD test. Fortunately for accountants and auditors forensically analyzing accounting and financial data, small amounts below \$100, \$50, or \$10 are not typically very important in the context of fraud detection, since that is not typically where fraud takes place. Rather, cheaters typically concoct numbers in the thousands and millions of dollars.

Alas, the entire discussion in this chapter is often superfluous, and usually the digital analyst can easily decide things based on the decimal precision point of the data under consideration. For dollar amounts with 2 decimal points of precision, all amounts below \$0.10 are excluded from higher orders, FTD and LTD. For example, \$0.07 contributes only to the first order. All amounts below \$1.00 are excluded from the third and higher orders as well as LTD. For example, the amount \$0.90 contributes 0 to second order, 90 to FTD, but not to LTD or third order. Only amounts over \$9.99 may contribute to LTD, and especially amounts over \$99.99. For example, the amount \$40.00 may contribute 00 to LTD, and 40 to FTD. For count data with 0 decimal points of precision, such as population data, a city with less than 10 inhabitants contributes only to the first order. A city with

less than 100 contributes only to first, second, and FTD tests, not to the third order or LTD. Only a city with more than 999 inhabitants, and especially over 9999 inhabitants, may contribute to LTD.

One should bear in mind that first-order test is not really sufficient for a thorough forensic analysis, and second-order, FTD, and LTD tests are an essential part of fraud detection. The chapter in the fourth section regarding the near indestructibility of higher-order-digit configurations suggests that higher-order tests should be considered by auditors and data analysts as by far more robust and reliable forensic tools than the mere application of the first-order distribution.

A better way to express much of what was said in this chapter is by simply acknowledging that even though the word “Last-Two-Digits” literally refers to the two digits on the rightmost side of the number, in reality the definition refers to the adjacent pair of digits belonging to the highest possible orders — and only if such a pair exists involving the third and fourth orders at a minimum.

## **Section 3**

### **DATA COMPLIANCE TESTS**

**This page intentionally left blank**

## TESTING DATA FOR CONFORMITY TO BENFORD'S LAW

---

The third section of this book attempts to answer the typical question asked by the anxious auditors, digital analysts, prosecution lawyers at governmental tax bureaus, and researchers once all digit distributions regarding the data set in question are obtained and clearly displayed, namely: “**Does the data obey Benford’s Law or not?**” The forensic digital analyst should keep in mind that no data set could ever conform exactly to Benford’s Law. Nor should we expect any such conformity when it comes to real-life random data. Hypothetically speaking, since any data set is finite, and since the law expresses the proportions as irrational numbers using logarithms, compliance is never perfect. The value  $\text{LOG}(1+1/d)$  is an irrational number that cannot be expressed as  $p/q$  with  $p$  and  $q$  being integers, therefore no actual ratio of (integral) numbers being led by digit  $d$  divided by the (integral) total number of entries in the data set exists that could ever be equal to such an expression. For example, assuming the data set in question contains 37,885 values, then for digit 1 the equality **(# of values led by digit 1)/(37,885) = log(2)** can never be true no matter how honest, natural, or supposedly logarithmic the data set is said to be, since  $\text{log}(2)$  is an irrational number. Even if a scheming and dishonest digital analyst hired by the tax authorities to conduct digital investigation would quietly whisper in the ears of the firm’s accountants prior to Benford’s Law compliance tests to quickly rewrite all their 37,885 numbers according to digits so as to appear as if they perfectly comply with the law, nothing would come out of it in terms of perfection. The corrupt digital analyst would hint at 11,404 or 11,405 numbers beginning with digit 1 as the most innocent-looking result for a data set having 37,885 values in all, but this leads to the ratios of  $(11404/37885)$  &  $(11405/37885)$  or 0.301016 & 0.301043, while  $\text{LOG}(1+1/1)$  equals 0.30102999566398 ... which is still stuck somewhere between these two ratios  $[0.301016 < \mathbf{0.301029} < 0.301043]$ , and no further improvement can be achieved here. Furthermore, tolerance of mild deviations from the law goes much deeper than the mere issue of the irrationality of the expressed probabilities; it has

to do with the fact that the law is merely a probabilistic one, as opposed to a deterministic law such as in physics or chemistry. Nobody visiting a casino ever gets really upset if in, say, 100 throws of a dice the lucky face of 6 showed up only 11 times instead of the expected 17 times [ $1/6 * 100 \approx 17$ ]. To immediately suspect the casino owner of using a highly sophisticated biased dice (where 6 is rarer than the usual  $1/6$  chance) would be considered highly unreasonable. A whole different conclusion though would be drawn about the casino if in, say, 100 throws of a dice the face of 6 showed up only twice. Analogically, no auditor should get upset if there are only, say, 28.5% numbers beginning with digit 1, yet, he or she should get quite suspicious if there are only say 14.7% such numbers. Where should we exactly place the cutoff points separating the reasonable from the suspicious, and conformity from deviation? Can it be based on some sound statistical theory, giving an exact and objective threshold over or below which results are considered significant at, say, 5% probability? The answer is unfortunately decisively negative. The harsh (statistical) reality is that the auditor or the researcher often needs to make a totally subjective decision regarding compliance! The reasons and rationale leading to this pessimistic conclusion shall now be explored.

Within the context of examining compliance with Benford's Law, there is a need to conceptually differentiate between two very different research questions, namely between **compliance** and **comparison**. The term compliance would refer to the case of a data set of a particular type, a type that is known to be generally logarithmic, while the auditor or the researcher is wondering about authenticity, errors, manipulations, or perhaps bad sampling. In other words, compliance is the quest of finding out whether the (relatively small) sample digitally follows the (relatively much larger) generic population logarithmic configuration. For example, for an auditor examining a particular revenue account of a single firm, the immediate and pressing question is whether to suspect fraud if digit configuration deviates somewhat from the logarithmic; given that generic revenue data is well-known to behave logarithmically. The question in this context is then: **"For this sample drawn from a supposedly logarithmic population, is digital deviation from the logarithmic due to chance or structural?"** The term comparison would refer to a particular processes, data set, or distribution that: (I) is not thought of as a sample and doesn't belong to any supposed larger population type or class, but rather standing alone in its own right, (II) is not thought of as logarithmic at all to begin with, (III) having no issue with honesty, error, or authenticity, (IV) having a particular digital configuration other than the logarithmic which is

believed to authentically belong to it. When such digital result is acknowledged, accepted, and respected in its own right, as it should be since Benford's Law does not apply to all real-life data sets; one legitimate question could still be asked in this context, namely: "**How far from the logarithmic is this digital configuration?**" One may inquire about the degree of deviation or a measure of 'distance' from the logarithmic. Even though the definition chosen to accomplish this — no matter how reasonable — would still be ultimately arbitrary, it is altogether fitting and proper that we should construct it.

To begin with, it is important to understand the underlying assumptions and background of the well-known chi-square test for compliance. The test is still widely and mistakenly used in the context of Benford's Law, causing many errors and misleading many scholars. The chi-square test refers to a sample of size  $N$ , randomly taken from an infinite population assumed to have the logarithmic property exactly. The inference and the focus here is not about the population's logarithmic property, which is taken as a given, but rather about the digital property and the integrity of the sample on hand. A large deviation from the logarithmic for large enough a sample is presumed to indicate an error, bad sampling procedure, or at the least a mishap of sorts, and possibly an evidence of intentional manipulation as in outright fraud.

Analyzing a piece of data digitally, with the intended purpose of comparing it to the logarithmic, implicitly implies that we are comparing this particular piece of data with some generic and larger **population** universe already known to be logarithmic. Say we have 57,000 revenue transactions from IBM relating to the first quarter of 2008. We are trying to investigate whether or not IBM gave honest information here, and so first-digit distribution is studied, compared to the logarithmic, and a decisive conclusion regarding honest/fraudulent reporting by IBM is arrived at via the chi-sqr test. Implicit in this whole forensic scheme is that the universe of revenue amounts, relating to all companies around the globe, obeys Benford's Law, which is indeed true. Surely, the hourly revenue of a tiny coffee shop on a street corner with a tiny clientele is not logarithmic, but the global aggregate — if the chief auditor at the UN or IMF could ever obtain such enormous and confidential data — is very nearly perfectly Benford. It is only in this context that the data on hand is considered a **sample**, a sample from a much larger population, and it is only in this context too that statistical theory can lend a hand and provides us with cutoff points and threshold values, by way of indicating their exact probabilistic significance. Yet, there is a serious pitfall in such an approach, namely that the nature of the sample data should resemble the nature of

the population in all its aspects, and this is rarely so, except in the mind of the eager, ambitious, and naïve auditor seeking an error-proof algorithm capable of detecting fraud where perhaps none exists. For example, the data on hand about IBM revenues is in reality the entire population, not a sample from that imaginary ‘universe of revenues’. This is so since each company has its own unique price list, particular clientele, unique products for sale, and belongs to some very specific sub-industry. Stated differently, the particular data set on hand, even if it can be thought of as a sample, was not ‘taken from the larger population’ in a truly random fashion. It is impossible to argue that these 57,000 revenue transactions from IBM database is a random sample from the universe of that generic and global revenue data type! To take a truly random sample from such imaginary universe/population, one should take say 25 values from IBM, 12 values from Nokia, 17 values from GM, 13 from Microsoft, and so forth, in which case results are guaranteed to be nearly perfectly logarithmic, and the application of the chi-sqr test wholly justified and workable!

An even more profound source of confusion and mistaken application of the chi-square test in the context of Benford’s Law is the fact that frequently it is impossible to contemplate or consider some larger data type standing as the population for the data set on hand. Frequently, what we have on hand is simply a unique set of numbers pertaining to a very particular issue or process, and attempting to envision it belonging to some ‘larger’ or ‘more-generic’ (and logarithmic) population data type is nothing but a fantasy, a misguided tendency to seek help from statistical theory where none exists. Even more incredible is to imagine the particular data set on hand as ‘a sample’ of sorts being generated in ‘a truly random fashion’ from that imaginary non-existent parental universe/population! As an example, suppose a researcher is investigating the particular phenomenon of chemical acidity in drinks. The researcher has worldwide data on all 12,658 known drinks, fruit juices, vegetable juices, alcoholic beverages, coffee, potions, and so forth, exhausting all possible sources, without neglecting a single exotic drink on some far away island or those drunk during rare religious and ceremonial occasions. The pH measure of all 12,658 drinks yields a particular first-digit distribution of, say, {41.8, 26.1, 5.0, 7.5, 2.4, 2.5, 3.4, 8.6, 2.7}. It does not make any statistical sense here to compare this digital configuration to the logarithmic via the chi-sqr distribution test (i.e. compliance). There is no null hypothesis to accept or reject. This pH data set is not a sample from some supposed larger universe/population to be compared with. Certainly the researcher cannot claim that his or her data on hand was obtained in a ‘truly random fashion’ from that imaginary larger

population of pH values! That would border on the absurd! This pH data set stands apart, proud and independent, existing in its own right. The issue of fraud does not enter here of course, and yet, one may legitimately wonder and ask “how far digital distribution of this pH data set is from the logarithmic” (i.e. comparison). There can't be, and there shouldn't be of course, any talk here of probabilistic 5% or 1% significance level, or of any supposed ‘chances’ of obtaining such a non-logarithmic pH digital result. This pH example is given for pedagogical purposes only; while a more realistic digit distribution for pH values would typically be perhaps as in: {0%, 3%, 21%, 13%, 37%, 19%, 7%, 0%, 0%}, where 2.0 is just about the most acidic drink, and 7.5 about the strongest alkali.

A totally different statistical scenario is demonstrated with the example of the mid-level or newly hired financial analyst or economist working at the IMF being asked by the top director of the institution to provide two large separate samples of revenues worldwide from a large mixture of companies for the years 2007 and 2008, as a comparison in the study of the causes leading to the onset of the Great Recession. On the (correct) assumption that global revenue data truly follows Benford's Law, the top director can discreetly perform a digital test on the data provided by the little known newly hired employee as a check whether (I) data was seriously and correctly collected in good faith, or that (II) workload was fraudulently reduced by simply concocting invented revenue numbers by the lazy and dishonest new analyst. In this example, the statistical methodology of the chi-square test (i.e. compliance) can and should be applied by the top director in order to detect fraud (or rather to detect laziness.)

To recap, the statistical backdrop of the following Z and chi-sqr tests is based on a truly random sample of size N taken from an infinitely large logarithmic population (or just sufficiently very large data set for all practical purposes). Here statistical theory can easily enter the fray and prescribe exact probability values and significant cutoff points for any digital configuration obtained for the sample.

There are two types of tests: (I) *An overall test* that takes into account all the digits or digit combinations by combining all nine (or more) deviations from expected proportions, (II) *A digit-by-digit test*, separately calculated for each digit or combination, where possibly some appear deviant and suspicious while others appear correct as expected. The relevant digits in question could be: 1 to 9 for the first order; 0 to 9 for the second order; 10 to 99 for the first-two-digits combination, and 00 to 99 for the last-two-digits combination. For forensic digital analysis in the context of Benford's Law, a significance level of 5% is more typical than 1%; yet this issue is ultimately a subjective decision.

## THE Z TEST

---

The Z Test is performed digit by digit or combination by combination and gives more specific information about which digits/combinations are off and which are not. Yet, Type I error (false positive) is much more likely for the Z test than for the overall test of the chi-square. Note the use of the letter **o** for “observed”, and the letter **e** for “expected”.

$P_e$  = the expected Benford proportion for the particular digit/combination.

$P_e = \text{LOG}(1+1/d)$  for the first-order test.

$P_e = \text{LOG}(1+1/pq)$  for the first-two-digits test (FTD).

$P_e = 1/100$  for the last-two-digits test (LTD).

$P_o$  = the observed actual proportion of the particular digit/combination in the data set.

$N$  = the number of values in the data set (# of observations).

$sd_d$  = the standard deviation of each digit/combination expected proportion.

$sd_d = \text{Square Root of } [P_e(1 - P_e)/N]$  for the digit/proportion in question.

Null Hypothesis  $H_0$ : Data obeys Benford’s Law in the context of the particular digit/combination. The Z test is performed individually on the particular digit/combination using the expression:

$$Z_d \text{ statistic} = (|P_o - P_e| - 1/(2N))/sd_d$$

An alternative expression for the Z statistic is:

$$Z_d \text{ statistic} = [(\sqrt{N}) * (|P_o - P_e| - 1/(2N))]/(\sqrt{(P_e(1 - P_e))})$$

Reject the Null Hypothesis  $H_0$  at the  $p\%$  confidence level if  $Z_d$  value is larger than  $Z$  with  $p\%$  critical value.  $Z$  refers to the Standardized Normal Distribution, namely the Normal with mean 0 and standard deviation 1. For example, the chance of being anywhere to the right of 1.960, or anywhere to the left of  $-1.960$ , on the Normal (0,1) is 5%. The chance of being anywhere to the right of 2.576, or anywhere to the left of  $-2.576$ , on the Normal (0,1) is just 1%. Critical values here are calculated as a two-tailed test. The term  $1/(2N)$  is called the ‘continuity

correction factor' and it is used only when its value is smaller than the absolute value term (hence guaranteeing that the  $Z_d$  statistic is never negative).

For example, for the first-digit set {34.8, 18.9, 13.4, 8.1, 5.7, 5.0, 4.8, 4.5, 4.7} of a sample of 1000 data points drawn from a very large logarithmic population, we get:

$$Z_1 \text{ statistic} = [(\sqrt{1000}) * (|0.348 - 0.301| - 1/2000)] / (\sqrt{(0.301(1 - 0.301))}) \\ = 3.18.$$

$$Z_2 \text{ statistic} = [(\sqrt{1000}) * (|0.189 - 0.176| - 1/2000)] / (\sqrt{(0.176(1 - 0.176))}) \\ = 1.03.$$

Since Z at the 5% confidence level is 1.960, it follows that the claim that the deviation of digit 1 was merely a chanced occurrence is statistically rejected at the 5% confidence level. On the other hand digit 2 does seem to comply with the law.

If the same first-digit set {34.8, 18.9, 13.4, 8.1, 5.7, 5.0, 4.8, 4.5, 4.7} is gotten with a much larger sample of 8000 data points, statistical theory is now strict and unforgiving, and it expects us to get digital results that are much closer to the logarithmic. Hence all digits 1–8 fail to pass the test, except for digit 9 which passes the test with the low  $Z_9$  statistic score of 0.51 due to its close proximity to the logarithmic value of 0.046.

$$Z_1 \text{ statistic} = [(\sqrt{8000}) * (|0.348 - 0.301| - 1/16000)] / (\sqrt{(0.301(1 - 0.301))}) \\ = 9.07.$$

$$Z_2 \text{ statistic} = [(\sqrt{8000}) * (|0.189 - 0.176| - 1/16000)] / (\sqrt{(0.176(1 - 0.176))}) \\ = 3.02.$$

$$Z_9 \text{ statistic} = [(\sqrt{8000}) * (|0.047 - 0.046| - 1/16000)] / (\sqrt{(0.046(1 - 0.046))}) \\ = 0.51.$$

For the quite deviant first-digit proportions {44.1, 17.7, 15.1, 9.7, 5.4, 3.1, 2.4, 1.5, 1.0} gotten from an extremely small sample of merely 30 data points, statistical theory is now highly forgiving and flexible and does not expect us to get results very close to the logarithmic (due to the low number of sample points). Here 'continuity correction factor' 1/60 or 0.0167 is larger than the absolute value term for digit 2 of  $|0.177 - 0.176|$  or 0.0010 and hence not incorporated into the statistic. Here, all digits 1–9 pass the test!

$$Z_1 \text{ statistic} = [(\sqrt{30}) * (|0.441 - 0.301| - 1/60)] / (\sqrt{(0.301(1 - 0.301))}) \\ = 1.472.$$

$$Z_2 \text{ statistic} = [(\sqrt{30}) * (|0.177 - 0.176|)] / (\sqrt{(0.176(1 - 0.176))}) = 0.013.$$

The Z test incorporates the term  $\sqrt{N}$ , implying that whenever data set is quite large, even a seemingly mild deviation from the logarithmic where  $|P_o - P_e|$  comes out quite small can still show a fairly large value of  $Z_d$  statistic, thus rejecting compliance with Benford's Law. Because of that the Z test is mistakenly thought to be "oversensitive", in the sense that for large data sets (say over 25,000 values) supposedly even mild deviations from Benford are flagged as significant ("false positive"). In other words, the test is erroneously thought to suffer from excessive power. This misguided perception of 'oversensitivity' is an indication that users lack understanding of the underlying statistical basis of the test. The dignified and well-respected statistician disguised as an addicted gambler and wearing some very casual clothes, confidentially sent to an ill-reputed casino by the authorities to investigate possible dishonesty and biasness of the large number of dice in use there, should certainly not be derogatorily referred to as 'oversensitive' or as 'overzealous' if he declares the casino to be fraudulent when after 10,000 throws of the dice only 1,027 times the lucky face of 6 showed up, instead of the expected 1,667 times [calculated as  $10000 \cdot (1/6)$ ]. The statistician should become increasingly overzealous and suspicious as the number of throws (N) increases and the ratio continues to significantly deviate from the theoretical  $1/6$  value. In any case, rightly or wrongly, the Z-test nowadays is used quite frequently whenever data set is relatively small, and it is erroneously avoided whenever the data set is deemed too large. Generally, auditors (mistakenly) consider any account with over 25,000 or 50,000 entries as too large for the Z-test. This is akin to the irrational critically ill patient asking the laboratory to return his blood tests only if results are negative and to discard the whole thing if it brings bad news. In reality, the Z test should usually be avoided altogether in the context of Benford's Law and regardless of data size, due to the often questionable basis of the underpinning statistical theory. Data size should never be the basis for deciding whether to apply the Z test or not, rather the correctness in the modeling of the data as a sample of some larger logarithmic population should be the only criteria of proper application.

## THE CHI-SQUARE TEST

---



---

The chi-square (chi-sqr) test is an overall test seeking to confirm Benford behavior for the data set as a whole regarding all relevant digits or digit combinations. Almost in all cases, the underlying theoretical and statistical basis for the chi-sqr test are not applicable to the data set under consideration, as seen by its supposed “oversensitivity” in the cases of large data sets where even mild deviations from the logarithmic are flagged as significant — as they should be! Yet, it has erroneously been used in accounting and auditing circles on a regular basis for many years, and unfortunately it is still being used nowadays as part of the standard procedure in fraud detection. This has led to much confusion and many errors, and has done a lot in general to undermine trust in the whole discipline of Benford’s Law. Even more unfortunate is its use in mathematical and empirical research where it is also erroneously applied blindly in almost all cases, lacking statistical justification, and has led to numerous misguided conclusions and much confusion.

Null Hypothesis  $H_0$ : The data set obeys Benford’s Law overall considering the set of all the digits/combinations in question, namely for (1 – 9), or (0 – 9), (10 – 99), or (00 – 99).

$$\text{chi-sqr statistic} = N * \sum_1^{\text{RD}} \frac{(\text{Pe} - \text{Po})^2}{\text{Pe}}$$

Where summation is run from the lowest possible digit or digit combination, to its maximum (called RD). That is, summed over (1 – 9), (0 – 9), (10 – 99), or (00 – 99).

Reject the Null Hypothesis  $H_0$  at the p% confidence level if chi-sqr value is larger than chi-sqr p% critical value with (RD – 1) degrees of freedom. RD is the number of Relevant Digits in the particular test. For example, RD = 9 for the first-order digits, RD = 10 for the second-order digits, RD = 90 for the first-two digits combination, and RD = 100 for the last-two digits combination. Note that rejecting the null hypothesis via the chi-sqr test does not tell us specifically which digits/combinations are problematic and which are not, which are overrepresented and which are underrepresented.

For the U.S. cities and towns population data of 19,509 centers which closely resembles the logarithmic having {29.4, 18.1, 12.0, 9.5, 8.0, 7.0, 6.0, 5.3, 4.6} as its first-digits proportions, a completely misguided application of the chi-sqr test here would yield the supposedly “chi-sqr statistic” of:

$$\begin{aligned} \text{chi-sqr}/N &= (0.294 - 0.301)^2/0.301 + (0.181 - 0.176)^2/0.176 + (0.120 - 0.125)^2/0.125 \\ &+ (0.095 - 0.097)^2/0.097 + (0.080 - 0.079)^2/0.079 + (0.070 - 0.067)^2/0.067 \\ &+ (0.060 - 0.058)^2/0.058 + (0.053 - 0.051)^2/0.051 + (0.046 - 0.046)^2/0.046 \\ &= 0.00016 + 0.00016 + 0.00019 + 0.00005 + 0.00001 + 0.00016 + 0.00005 \\ &+ 0.00010 + 0.00001 = \mathbf{0.000898}. \end{aligned}$$

Hence chi-sqr/N is 0.000898, and chi-sqr is 0.000898\*N = 0.000898\*19509 = 17.36 with eight degrees of freedom. The calculations above carried more precision than what is apparent here, as nine decimal places values of proportion of population and BL were carefully carried. In the table for eight-d.o.f. chi-sqr, the cutoff point for significance level of 5% is 15.51. Since our 17.36 “chi-sqr statistic” is higher than 15.51, supposedly then the null hypothesis that this population data is Benford must be rejected! This chi-sqr rejection of our nicely logarithmic data set is not a repudiation of the whole science of statistics and mathematics, but rather simply the misguided use of it, applying it where it should not be applied. U.S. population data is simply not a truly random sample from some supposedly larger imaginary universe of logarithmic data. Had this first-digits proportion been gotten from say 19,509 truly random picks from our Aggregate Global Data via a methodical gathering of numbers from a huge variety of sources frequently shifting the source, say, each 3 values, then such chi-sqr application would be wholly justified, and all its conclusions correct and accepted. Statistical theory simply demands more from such enormous (large N) random pick from AGD, and it is certain that 95 out of 100 such undertakings would be confirmed by 5% confidence level chi-sqr test, and that the researcher would end up 95% of the times with digital configurations that are even closer to the logarithmic than the one gotten for the U.S. population data.

The supposedly ‘troublesome’ value of N is noted within the chi-sqr expression above as a multiplicative factor, implying that for any given magnitude of deviation (actual from expected) the statistics still depends on N and increases accordingly. When N is quite large the test seems to become ‘too sensitive’ in the eye of the non-statistician, and bitter complaints about the N term within the

algebraic expression are frequently heard, as even tiny deviations from the Benford proportion flag the data set as significantly non-logarithmic. While such oversensitivity is perfectly proper and statistically correct if test is applied under the right circumstances, lack of statistical understanding has caused many to misguidedly call it ‘false positive’ and to claim that the chi-sqr test itself suffers from ‘excess power’. Yet, when the underlying basis of applying the chi-sqr is valid, namely that the data was drawn in a truly random fashion from a truly logarithmic population, the chance of finding deviations from the logarithmic in our sample (data on hand) is closely related to data size  $N$ , and probability of even minute deviation sharply diminishes whenever size is large. Analogously, our statistician turned detective investigating supposed fraud by the casino owners would accept not seeing a single face of six in 10 throws, but not if the dice is thrown 1000 times all to no avail without a single face of six. This is simply the consequence of the law of large numbers. Certainly, the statistician would still certify the casino as fraudulent even if the face of six pops out say 50 times in 1000 throws (5%), but wouldn’t be suspicious at all if out of 10 throws six never pops out (0%)! As the number of dice trials increases, we demand better accuracy with that supposed  $1/6$  probability value (16.6%), hence for only 10 throws the low value of 0% is quite acceptable, while for 1000 throws not even 5% is enough for establishing trust, and the unbiased-ness of the dice is called into serious doubt.

In order to highlight the supposed difficulties in applying the chi-sqr test, and to demonstrate the typical complaints against the test, two single-issue physical data sets encountered earlier would be compared. The first data set relates to the length of 158 major rivers in Canada. The size of this data set is very small, and its first-order digit distribution is  $\{21.5, 17.1, 14.6, 15.8, 9.5, 6.3, 6.3, 5.1, 3.8\}$ . The second set (provided by the Australian National University in Canberra) relates to data on the time interval in seconds between 2,258,653 consecutive earthquakes worldwide in the period 01/01/1970 to 12/31/2009, with no restrictions on geographical position, depth, or magnitude. The size of this data set is truly huge, well in excess of two million values! First-digit distribution here is  $\{29.1, 17.2, 12.6, 10.0, 8.2, 6.9, 5.9, 5.2, 4.6\}$ .

Researchers would instinctively rush now to compare these two results by the misguided examinations of the chi-sqr statistics. The data on rivers is clearly quite different from Benford, yet due to its extremely small size it easily passes the test at the 5% confidence level! Its chi-sqr statistic is evaluated at 11.37, which

is comfortably below the 15.51 cutoff point of 5% for the eight-d.o.f. chi-sqr distribution. The data on earthquakes on the other hand, which is remarkably so close to the logarithmic, cannot pass the chi-sqr test in the least, as its chi-sqr statistic is evaluated at 1758.67, well above the threshold of 15.51 cutoff point of 5% for the eight-d.o.f. chi-sqr distribution. This result and comparison is a bit shocking at first, yet when one contemplates the severe mistake committed in applying here the chi-sqr test in the first place, this 'paradoxical' result becomes quite clear. Data in neither case was drawn from any supposed larger population of logarithmic data, and certainly not in any random fashion from any imaginary universe of such data. The chi-sqr test cannot and should not have been used here for these two data sets to begin with!

As an exercise in the usage of the chi-sqr test with higher degrees of freedom, an imaginary example of second-order-digit distribution of a truly random sample of 1200 values from AGD shall be examined. Assuming that the distribution is {13.3, 12.0, 11.3, 11.0, 10.3, 10.2, 9.4, 9.2, 7.1, 6.2}, it shall be compared to the logarithmic {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5} via the chi-sqr statistics.

$$\begin{aligned} \text{chi-sqr}/N &= (0.133 - 0.120)^2 / 0.120 + (0.120 - 0.114)^2 / 0.114 + (0.113 - 0.109)^2 / 0.109 \\ &+ (0.110 - 0.104)^2 / 0.104 + (0.103 - 0.100)^2 / 0.100 + (0.102 - 0.097)^2 / 0.097 \\ &+ (0.094 - 0.093)^2 / 0.093 + (0.092 - 0.090)^2 / 0.090 + (0.071 - 0.088)^2 / 0.088 \\ &+ (0.062 - 0.085)^2 / 0.085 = 0.00141 + 0.00032 + 0.00015 + 0.00035 + 0.00009 \\ &+ 0.00026 + 0.00001 + 0.00004 + 0.00328 + 0.00622 = \mathbf{0.01213}. \end{aligned}$$

Hence  $\text{chi-sqr}/N = 0.01213$ , and  $\text{chi-sqr} = 0.01213 * N = 0.01213 * 1200 = 14.6$ .

In the table for nine-d.o.f. chi-sqr, the cutoff point for significance level of 5% is 16.9. Since this 14.6 chi-sqr statistic is lower than 16.9 the null hypothesis that this data is Benford in the second order cannot be rejected. Rather it is accepted at the 5% confidence level as being logarithmic.

An equivalent expression for the chi-sqr statistic utilizes the definition  $P_o = O_d / N$  where  $O_d$  represents the actual observed number of values of digit  $d$  or digit combination, as well as the definition  $P_e = B_d / N$  where  $B_d$  represents the expected number of values of digit  $d$  or digit combination according to Benford's Law (given  $N$  values in the data set), hence:

$$\begin{aligned} \text{Chi-sqr statistic} &= N * \sum [(P_e - P_o)^2 / P_e] \\ \text{Chi-sqr statistic} &= N * \sum [(B_d / N - O_d / N)^2 / (B_d / N)] \end{aligned}$$

$$\text{Chi-sqr statistic} = N * \sum [(1/N)^2 (B_d - O_d)^2 / (B_d/N)]$$

$$\text{Chi-sqr statistic} = \sum (B_d - O_d)^2 / B_d.$$

where quantities  $B_d$  and  $O_d$  represent the theoretical and actual number of values of digit  $d$  or digit combination within the data set, not proportions.

## SSD AS A MEASURE OF DISTANCE FROM THE LOGARITHMIC

---

Let us construct a measure of ‘distance’ from the actual/observed digital proportions of any data set to the ideal Benford’s digital proportions, namely a measure of deviation from the logarithmic expectation. It must be emphasized here that we are not seeking to measure compliance or conformity with Benford’s Law; neither do we attempt here to statistically decide on whether a given deviation is significantly or not significantly different from Benford. Testing for compliance with Benford’s Law and calculating a comparison measure of distance from the logarithmic for a given digital configuration are two separate things. The **Sum Squares Deviation (SSD)** comparison measure suggested here is one that is independent of the number of observations  $N$ , and as such it cannot give rise to any statistical theory to guide us as to significant values for rejecting or accepting compliance with the law (exact threshold or cutoff points). SSD measure can only be used as a valid comparison with the logarithmic for all data types, large-sized data sets or small-sized ones, samples or whole populations. The definition of SSD is the identical twin of SSE (sum of squares error) in regression theory and general residual analysis. The definition of SSD does not compare fractions or proportions (e.g. **0.301**), but rather percent (e.g. **30.1**). This is so in order not to deal with extremely small fractional values which are often confusing and very hard to remember.

$$\text{Sum Squares Deviation (SSD)} = \sum_1^{\text{RD}} (T_o - T_e)^2$$

Where summation is run from the lowest possible digit or digit combination, to its RD maximum. That is, summed over (1 – 9), (0 – 9), (10 – 99), or (00 – 99).

$T_o$  = the **observed** actual **percent** of numbers of digit  $d$  or combination.

$T_e$  = the **expected** Benford **percent** of numbers of the digit or combination.

$T_e = 100 * \text{LOG}(1 + 1/d)$  for the first order.

$T_e = \{12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5\}$  for the unconditional second order.

$T_e = 100 * \text{LOG}(1+1/pq)$  for the first-two-digits combination (FTD).

$T_e = 100 * (1/100)$ , or simply 1, for the last-two-digits combination (LTD).

For the time between 2012 earthquakes data set in units of seconds, with first digits {29.9, 18.8, 13.5, 9.3, 7.5, 6.2, 5.8, 4.8, 4.2}, SSD measure of distance from the logarithmic is calculated as:

$$\text{SSD} = (29.9 - 30.1)^2 + (18.8 - 17.6)^2 + (13.5 - 12.5)^2 + (9.3 - 9.7)^2 + (7.5 - 7.9)^2 + (6.2 - 6.7)^2 + (5.8 - 5.8)^2 + (4.8 - 5.1)^2 + (4.2 - 4.6)^2 = 3.1$$

For the U.S. cities and towns population data which also closely resembles the logarithmic with first digits {29.4, 18.1, 12.0, 9.5, 8.0, 7.0, 6.0, 5.3, 4.6}, SSD measure of distance from the logarithmic is calculated as:

$$\text{SSD} = (29.4 - 30.1)^2 + (18.1 - 17.6)^2 + (12.0 - 12.5)^2 + (9.5 - 9.7)^2 + (8.0 - 7.9)^2 + (7.0 - 6.7)^2 + (6.0 - 5.8)^2 + (5.3 - 5.1)^2 + (4.6 - 4.6)^2 = 1.3$$

Hence U.S. population centers data might be thought of as being 'closer to the logarithmic' than the data on time between earthquakes — given SSD measure of 'distance'.

For the data on river length in Canada which is quite different from the logarithmic with first digits {21.5, 17.1, 14.6, 15.8, 9.5, 6.3, 6.3, 5.1, 3.8}, SSD measure of distance from the logarithmic is calculated as:

$$\text{SSD} = (21.5 - 30.1)^2 + (17.1 - 17.6)^2 + (14.6 - 12.5)^2 + (15.8 - 9.7)^2 + (9.5 - 7.9)^2 + (6.3 - 6.7)^2 + (6.3 - 5.8)^2 + (5.1 - 5.1)^2 + (3.8 - 4.6)^2 = 119.5$$

The table in Fig. 3.1 shows a variety of real and imaginary digital configurations, together with their corresponding SSD measures.

For second-order digit distribution, SSD sums up 10 squared terms. For FTD distribution, SSD sums up 90 squared terms of typically much lower values. For LTD distribution, SSD sums up 100 squared terms of such low values.

For U.S. Census cities and towns population data, second-order leading digits proportions are {11.9, 11.4, 11.3, 10.5, 10.2, 9.4, 9.6, 8.7, 8.8, 8.1}, hence its SSD measure of distance from the logarithmic in relation to the unconditional second order of {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5} is calculated as:

$$\text{SSD} = (11.9 - 12.0)^2 + (11.4 - 11.4)^2 + (11.3 - 10.9)^2 + (10.5 - 10.4)^2 + (10.2 - 10.0)^2 + (9.4 - 9.7)^2 + (9.6 - 9.3)^2 + (8.7 - 9.0)^2 + (8.8 - 8.8)^2 + (8.1 - 8.5)^2 = 0.6$$

DIGIT	US Population Cities & Towns	Time Between Earth-quakes	Earthquake Depths	Global Infectious Disease Cases	Common Chemical Compounds	River Length Canada	Imaginary Highly Skewed Digits	US Census County Area	Imaginary Inverted Digits
1	29.4	29.9	31.6	33.7	31.9	21.5	41.0	16.2	6.9
2	18.1	18.8	16.9	16.7	25.2	17.1	23.5	10.0	7.5
3	12.0	13.5	14.0	13.2	16.1	14.6	11.0	10.7	8.5
4	9.5	9.3	8.7	10.7	8.4	15.8	6.0	15.8	10.4
5	8.0	7.5	7.0	7.3	5.7	9.5	5.0	15.2	11.2
6	7.0	6.2	7.4	5.4	4.3	6.3	4.5	10.4	11.5
7	6.0	5.8	5.3	4.6	2.9	6.3	4.0	8.6	13.0
8	5.3	4.8	4.6	5.1	3.2	5.1	2.0	7.1	14.0
9	4.6	4.2	4.4	3.3	2.3	3.8	3.0	5.9	17.0
SSD	1.3	3.1	8.0	20.4	102.9	119.5	198.1	372.0	976.0

Figure 3.1 SSD Measure of the 1st Digits for a Variety of Digital Configurations

For SSD calculations of first-two-digits (FTD) proportions for US population data, the first three FTD proportions of 10, 11, 12, and those of the last three combinations 97, 98, 99 are {4.1%, 3.6%, 3.5%, and so forth, ..., 0.42%, 0.51%, 0.41%}. FTD in Benford’s Law is {4.1%, 3.8%, 3.5%, and so forth, ..., 0.45%, 0.44%, 0.44%}, hence SSD is calculated as:

$$SSD = (4.1 - 4.1)^2 + (3.6 - 3.8)^2 + (3.5 - 3.5)^2 + \text{and so forth, } \dots, + (0.42 - 0.45)^2 + (0.51 - 0.44)^2 + (0.41 - 0.44)^2 = 0.5$$

Prior to the calculation of last-two-digits (LTD) proportions for U.S. population centers data, those cities and town with less than 1,000 inhabitants should be eliminated from the data, resulting in a data set having only 10,215 cities. This guarantees that LTD proportion truly represents approximately the third and fourth orders — or some higher order. The first three LTD proportions of 00, 01, 02, and those of the last three combinations 97, 98, 99 are {1.11%, 1.08%, 1.01%, and so forth, ..., 0.93%, 0.83%, 0.92%}, hence SSD is calculated as:

$$SSD = (1.11 - 1.00)^2 + (1.08 - 1.00)^2 + (1.01 - 1.00)^2 + \text{and so forth, } \dots, + (0.93 - 1.00)^2 + (0.83 - 1.00)^2 + (0.92 - 1.00)^2 = 0.9$$

Let us now express SSD in a slightly different form in order to compare it with the expression of the chi-square statistic. Since  $T_o = 100 * P_o$  and  $T_e = 100 * P_e$ , we get:

$$\begin{aligned} \text{Sum Squares Deviation (SSD)} &= \sum (100 * P_o - 100 * P_e)^2 = \sum 100^2 * (P_o - P_e)^2 \\ \text{Sum Squares Deviation (SSD)} &= 10000 * \sum (P_o - P_e)^2 \end{aligned}$$

$P_e$  = the expected Benford proportion for the particular digit/combination.

$P_o$  = the observed actual proportion of the particular digit/combination.

This last form of SSD appears quite similar to the definition of the chi-sqr statistic, except for two differences, namely the absence of that supposedly 'problematic'  $N$  value in the expression, as well as not having to divide the main part  $(P_o - P_e)^2$  by the proportion  $P_e$ .

Since SSD does not incorporate the term  $N$  in its expression, one gets the same measure and conclusion regardless of the number of observations in the data set. This comes at a certain price as there is no associated statistical theory to guide us. There is also no hope that future statistical studies would somehow yield threshold points, significant values, or confidence intervals by applying SSD, since those highly beneficial results would certainly require involving  $N$  somewhere in the relevant expression, while  $N$  is nowhere to be found in the definition of SSD. Statistical theory cannot be indifferent to data size  $N$ , since there is no way to tell whether a certain deviation from the logarithmic is due to chance or to structural causes without knowing how many values have been collected as a sample from that supposedly larger logarithmic population. SSD is therefore applied only as a measure of distance from the logarithmic, and one has to **subjectively** judge a given SSD value of the data on hand to be either small enough and thus somewhat close to the logarithmic, or too high and definitely non-logarithmic in nature. A more systematic way of going about it is to empirically compare SSD of the data set under consideration to the list of a large variety of SSD values of other honest and relevant data sets of the same or similar type (i.e. belonging to the same topic). Such accumulated knowledge helps us in empirically deciding (in a non-statistical and non-theoretical way) on the implications of those SSD values. Yet some subjectivity is unfortunately necessary here in choosing cutoff points. The table in Fig. 3.2 is just one such summary list of SSD performed for the first-order, second-order, FTD, and LTD distributions for a variety of data sets and distributions. Some of the data sets in the table shall be explored in detail in later chapters. Almost all those that are extremely close to being logarithmic (the ones in the first and second groups) come with four SSD measures that are all less than two. Also of note here is that for the last group in the list, second-order in general came out quite close to the logarithmic (low SSD) even within data sets having first-order distribution quite far from the logarithmic (high SSD). In other words, second-order is found

<b>Data Set:</b>	<b>SSD -----&gt;</b>	<b>1st</b>	<b>2nd</b>	<b>FTD</b>	<b>LTD</b>	<b>Data Points</b>
=====	=====	=====	=====	=====	=====	=====
<b>U.S Population Centers</b>		<b>1.3</b>	<b>0.6</b>	<b>0.5</b>	<b>0.9</b>	19509
<b>Earthquake (time between)</b>		<b>3.1</b>	<b>0.7</b>	<b>0.8</b>	<b>7.1</b>	19451
<b>Mixed 34,269 Data Points (Hill)</b>		<b>4.8</b>	<b>0.5</b>	<b>6.7</b>	<b>9.4</b>	34268
<b>UUUUU(0,10000) - 5 Uniforms Chain</b>		<b>0.1</b>	<b>0.6</b>	<b>0.6</b>	<b>0.5</b>	20072
<b>exponential(U(0, U(0, 13))) - Chain</b>		<b>1.7</b>	<b>0.1</b>	<b>0.3</b>	<b>0.3</b>	39995
<b>Lognormal, Location=9.3, Shape=1.7</b>		<b>0.2</b>	<b>0.3</b>	<b>0.3</b>	<b>0.3</b>	34999
<b>Stars Distance from Solar System</b>		<b>14.1</b>	<b>0.4</b>	<b>2.5</b>	<b>9.3</b>	48110
<b>Pulsar Frequency Data</b>		<b>37.1</b>	<b>4.7</b>	<b>6.8</b>	<b>5.7</b>	2208
<b>Exoplanet Period</b>		<b>65.0</b>	<b>4.7</b>	<b>15.7</b>	<b>23.5</b>	800
<b>Exoplanet Mass</b>		<b>10.7</b>	<b>30.9</b>	<b>12.5</b>	<b>26.0</b>	800
<b>Carbon Dioxide Emission</b>		<b>62.6</b>	<b>47.2</b>	<b>49.9</b>	<b>53.0</b>	217
<b>Chemical Compounds (Molar Mass)</b>		<b>102.9</b>	<b>5.3</b>	<b>19.8</b>	<b>66.5</b>	2175
<b>Oklahoma ST Payroll - 1st Q, 2012</b>		<b>112.5</b>	<b>22.3</b>	<b>31.3</b>	<b>95.5</b>	19675
<b>U.S County Area</b>		<b>371.8</b>	<b>7.7</b>	<b>46.0</b>	<b>14.7</b>	3143
<b>Periodic Table (Atomic Weight)</b>		<b>534.1</b>	<b>87.5</b>	<b>107.6</b>	<b>90.7</b>	117

Figure 3.2 SSD Measures on First, Second, FTD, and LTD for a Variety of Data Sets

in general to be much closer to the logarithmic as compared with first-order! This empirical finding is nicely consistent with what shall be seen and discussed in a later chapter in the fourth theoretical section on the near indestructibility of higher-order distributions.

Let us provide those arbitrary cutoff points as criteria. For the first digits, SSD generally should be below 25; those over 100 are considered to deviate too far from the logarithmic; and a reading below 2 is considered to be ideally Benford. For the second-order as well as for the first-two-digits, SSD should be below 10; a reading over 50 is considered non-logarithmic for the most part; while a reading below 2 is considered to be ideally Benford. For the last two digits, SSD should be below 40; a reading over 100 is considered non-logarithmic; while a reading below 4 is considered to be ideally Benford. LTD test is given a bit more slack since [among other reasons] often the last-two-digits are not available for all the data points as truly equal in distribution. The table in Fig. 3.3 summarizes the criteria. These rough guidelines could be further refined to some more specific data types, to zoom in on particular classes of data such as revenue in accounting, number of

<b>SSD</b>	<b>≈ Perfectly Benford</b>	<b>Acceptably Close</b>	<b>Marginally Benford</b>	<b>Non-Benford</b>
<b>1st Order</b>	< 2	2 - 25	25 - 100	> 100
<b>2nd Order</b>	< 2	2 - 10	10 - 50	> 50
<b>First-Two-Digits</b>	< 2	2 - 10	10 - 50	> 50
<b>Last-Two-Digits</b>	< 4	4 - 40	40 - 100	> 100

**Figure 3.3** Arbitrary SSD Cutoff Points for First, Second, FTD, and LTD Distributions

accidents per city in census data, infectious disease cases in biological data, and so forth.

It is not unreasonable to perform an inter-order comparison of SSD values within a given data set, and the relative stability of the second order as compared with the first order was discussed earlier. Even though the definition of SSD does not involve the concept of [average] deviation per digit or per digit combination, while the four tests come with varying numbers of digits or digit combinations, they still could end up about equal. For example, for the Lognormal distribution with 1.7 shape parameter and 9.3 location parameter, SSD values for the set of all four tests are  $\{0.2, 0.3, 0.3, 0.3\}$ , which are almost identical. The main explanation for this is that for the first order, for example, there are only nine such squared differences, yet typically variations there are high around the high-valued set  $\{30.1\%, 17.6\%, \dots, 4.6\%\}$ , while for the last-two-digits test, for example, there are 100 such values but variations there are low, fluctuating tightly around the  $\{1\%, \dots, 1\%\}$  low theoretical proportion.

## SAVILLE REGRESSION MEASURE

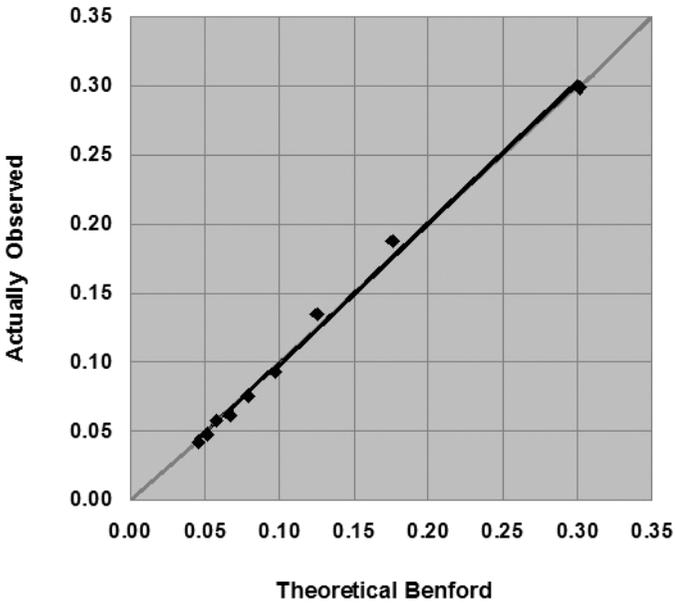
---

---

A novel approach in assessing the relationship between **first** leading digits of a given data set and the Benford proportions of  $\text{LOG}(1+1/d)$  was suggested by Adrian Saville in 2006. The algorithm is independent of the data size  $N$ , and only the resultant digital proportions of the data set are considered and compared with those of the Benford proportions.

Saville's algorithm is based on applying the classic simple linear regression model  $Y = \mathbf{b} * X + \mathbf{a}$  to **Observed** =  $\mathbf{b} * \text{Benford} + \mathbf{a}$ , where  $X$ , the independent variable, represents the theoretical Benford proportions, and  $Y$ , the dependent variable, represents the actual observed proportions for the data set under consideration. Parameter  $\mathbf{b}$  is the slope and parameter  $\mathbf{a}$  is the intercept. Although Saville's algorithm is erroneously presented as a compliance test, in reality it cannot serve as such given the total absence of the data size  $N$  in the definition of the statistic, a value that is indispensable in the theoretical construction of any statistical model yielding significant cutoff points or confidence intervals. Rather, the algorithm is essentially a comparison measure, focusing on the closeness of resultant slope  $\mathbf{b}$  to the ideal value of 1. For data sets that are nearly logarithmic, resultant regression slope  $\mathbf{b}$  should be very close to 1. Resultant slope value much larger or much smaller than 1 is an indication that the data set is not logarithmic (and thus possibly fraudulent if it represents accounting or financial amounts).

Let us illustrate Saville's method by applying it to the geological data set on time in seconds between successive earthquakes for the entire year of 2012. The first-digit proportions of this earthquake data are designated  $Y$ , and it is regressed on the Benford proportions  $\text{LOG}(1+1/d)$  designated as  $X$ . Figure 3.4 depicts Saville's scatter plot of the observed nine points together with the resultant regression line. The gray  $45^\circ$  line of  $Y = X$  is the ideal Benford line which a perfectly logarithmic data set generates. The black line is the simple linear regression line of the actual data points with slope  $\mathbf{b}$  and intercept  $\mathbf{a}$ . Certainly the issue of fraud does not enter here since Mother Nature is well-known for her honesty, even when she ferociously shakes the Earth in her rare moments of anger. The algorithm simply measures the closeness of the earthquake data to Benford without being at all judgmental about her actions.



**Figure 3.4** Saville Regression Plot for Time between 2012 Earthquakes Data Set

Earthquake first digits are {29.9, 18.8, 13.5, 9.3, 7.5, 6.2, 5.8, 4.8, 4.2}.

Saville slope = **1.02** Saville intercept =  $-0.0026$  **nearly Benford**

The closeness of the earthquake data to the logarithmic can be visualized by way of the closeness of the black regression line to the ideal Benford gray  $45^{\circ}$  line. The two lines almost overlap, signaling strong similarity to the logarithmic. The slope value of 1.02 is extremely close to the ideal Benford slope of 1, confirming the near-perfect logarithmic behavior.

Let us apply Saville's method to the geological data set on Geomagnetic Reversals seen in Fig. 1.23. Figure 3.5 depicts Saville's scatter plot for this data set together with the black regression line which is steeper than the ideal Benford gray line.

Geomagnetic Reversals first digits are {32.3, 19.4, 13.9, 11.8, 5.3, 4.3, 3.2, 5.4, 4.3}.

Saville slope = **1.16** Saville intercept =  $-0.0175$  **skewer than Benford**

Some deviation from the logarithmic is seen here. The data set can be judged to be skewer than the logarithmic and in favor of low digits over and above their usual advantage under the law of Benford. This can be visually seen in the scatter plot, and which is also confirmed by the value of the slope being larger than 1.

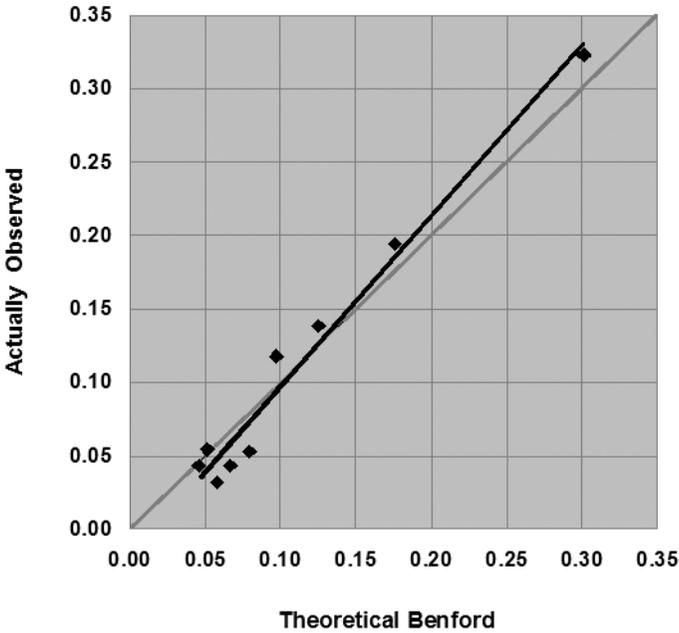


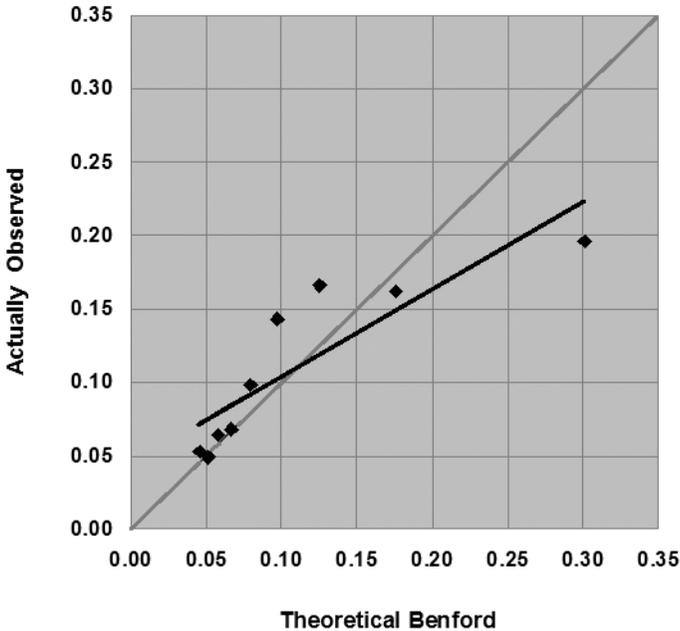
Figure 3.5 Saville Regression Plot for Geomagnetic Reversal Data Set

Let us apply Saville’s method to the monthly total returns of Athena Guaranteed Futures Ltd Fund discussed earlier in the previous section and depicted in Fig. 2.7. Figure 3.6 depicts Saville’s scatter plot together with the black regression line which is flatter (less skewed) than the ideal Benford gray line.

Athena Fund first digits are {19.6, 16.2, 16.6, 14.3, 9.8, 6.8, 6.4, 4.9, 5.3}  
 Saville slope = **0.59** Saville intercept = 0.0454 **less skewed than Benford**

Substantial deviation from the logarithmic is seen here. The data set can be judged to be less skewed than the logarithmic overall, as visually seen in the scatter plot and confirmed by the value of the slope being much less than 1.

Closeness to Benford in this context necessitates that the slope **b** be very close to 1, and/or that the intercept **a** would be very close to 0, although our focus shall be given exclusively to the value of the slope, ignoring the intercept altogether. How far from 1 should the slope be in order for the data set to be considered too deviant from the logarithmic, not complying with the law? Are there any clear cut-off/threshold points to guide us? Unfortunately there exists no such decisive rule, given the total absence of the data size N in the definition of the slope statistic, just



**Figure 3.6** Saville Regression Plot for Athena Guaranteed Futures Ltd Fund Data

as in the SSD case. Instead of an objective criterion as to conformity, only subjective judgment as to closeness may be exercised here. Moreover, a perfect slope of 1 does not necessarily imply that the data itself is logarithmic, since these 9 points could be envisioned as scattering wildly around their regression line and greatly deviating from the logarithmic. Yet, even though Saville's algorithm cannot serve even as a reliable comparison measure, Saville's slope turned out to be an elegant and concise measure of skewness relative to the normal degree of skewness of the logarithmic condition. A slope value over 1 indicates that digits are skewer than the Benford condition in favor of low ones. A slope value less than 1 indicates that digits are less skewed as compared with the Benford condition. Such convenient measure of relative skewness turned out to be immensely useful in precisely measuring digital development pattern, which is essentially all about shifts in skewness along the entire range of data.

A separate chapter with detailed analysis of the theoretical basis and all possible applications of Saville method is given in Section 6.

## VALUE REPETITION TEST

---

---

Value Repetition is a non-digital complementary test essential in forensic analysis for accounting and financial data in the context of fraud detection. It is performed by examining those repetitions of actual values (not digits) within the data set occurring relatively all too frequently. A repetition is an amount occurring within the data more than once. For example, if \$3,570 has occurred seven times, then it draws the attention of the auditor and may possibly lead to some source of actual fraud. If \$3,280 has occurred 157 times, then this amount is a prime candidate for possible fraudulent activity in the account, and must be thoroughly examined or explained. Could it be that \$3,280 was simply concocted 157 times in the imaginative mind of the cheating accountant (and thus contributing to spikes of 3, 2, 32, 80 on the first-order, second-order, FTD, LTD charts respectively)? Value repetition test is done in conjunction with digital analysis since repetitions may help in identifying the specific amounts that were causing the spikes on the charts. To perform this test a table is created showing all repeated values together with their associated frequency of occurrence (typically done in MS Access, other database software, or even in MS Excel spreadsheet). These two columns are normally sorted (high to low) by frequency and presented in decreasing order for easy inspection. It is important to eliminate the rest of the (normally) huge table below the top so as to focus only on those amounts that truly occur extremely too frequently. We are not really interested in amounts that occurred twice or three times, but rather in those occurring most frequently.

An example of Value Repetition Test is depicted in the table shown in Fig. 3.7 representing revenue amounts for a hypothetical company. The high frequency of amounts \$7.30, \$4.37, and \$549.00 might be explained away if the company has some very popular items for sale costing exactly those amounts, or half/third/quarter of those amounts. If there is a popular item on sale costing \$3.65 (half of \$7.30), then a client purchasing two items would generate revenue in the amount of \$7.30, and this perhaps could explain the high frequency of \$7.30 occurrences.

<b>Amount</b>	<b>Frequency</b>
<b>\$ 7.30</b>	<b>296</b>
<b>\$ 4.37</b>	<b>175</b>
<b>\$ 549.00</b>	<b>134</b>
<b>\$ 350.00</b>	<b>77</b>
<b>\$ 75.00</b>	<b>76</b>
<b>\$ 19.95</b>	<b>74</b>
<b>\$ 195.00</b>	<b>50</b>
<b>\$ 100.00</b>	<b>48</b>
<b>\$ 60.00</b>	<b>47</b>
<b>\$ 6.62</b>	<b>42</b>
<b>\$ 250.00</b>	<b>39</b>
<b>\$ 12.25</b>	<b>35</b>
<b>\$ 175.00</b>	<b>35</b>
<b>\$ 50.00</b>	<b>35</b>
<b>\$ 99.00</b>	<b>33</b>
<b>\$ 19.75</b>	<b>33</b>

**Figure 3.7** Value Repetition Test

As mentioned earlier, it is difficult for people to act truly randomly even when they are asked to do so, and they tend to repeat particular numbers that they subconsciously favor. Even statistics-major students or casino employees fumble when asked to write down 50 imaginary throws of a dice, giving quite unnatural, distinctly non-uniform and rare set of 50 numbers from  $\{1 \text{ to } 6\}$ . Fraudulent repetitions in real-life accounting data occur mostly when the fake data spans an elongated period of time, representing numerous number-invention occasions, when the cheater does not remember his or her manner of cheating in the past, and simply repeats inventing the same types of numbers all over again. Value Repetition Test can be quite useful in fraud detection, as it can possibly point to those fake numbers that the repeated cheater subconsciously favors.

As mentioned earlier, excessive repetitions of amounts in and of themselves do not prove that fraud has been committed, since the particulars of the price list of the company and the purchasing habits of its clients can cause that. For example, if an apparel shop prices its most popular jeans at \$32.80, then 3 would be over-represented in the first-digit order of its revenue data, and 80 in the last-two-digits distribution. Since a purchase of two pairs of these jeans costs \$65.60, then to a

lesser degree 6 would also be overrepresented in the first-digit order of its revenue data, and 60 in the last-two-digits distribution. Durtschi *et al* (2004) suggest eliminating from the accounting data under examination all those repeated entries peculiar to the operations of the company under consideration prior to any digital forensic analysis, and to focus only on digital results from the rest of the data. An example of such an analysis on a large medical center can be found on page 23 in their article. Dorrell and Gadawski (2012) suggest adding a third column showing the percent of total disbursement in order to direct more attention to those repeated transactions involving the highest aggregate amounts.

## THE CONFUSION AND MISTAKEN APPLICATIONS OF SUMMATION TEST

---

---

Worse than the misapplication and confusion regarding the chi-sqr test, Summation Test stands out as one of the most misguided application in the whole field of Benford's Law, attaining recently the infamous status of a fictitious dogma and leading many accounting departments and tax authorities astray. Erroneous applications aside, the test itself is theoretically fascinating as it relates to actual amounts in conjunction with digital leadership, as opposed to either the traditional approach of examining data solely as amounts via descriptive statistics or the latest Benfordian machinery of examining only digital distributions separately. Classic Benford analyses throw all the numbers of a given data set into nine distinct bins, count the numbers falling within each bin, and expect to get about  $100 \cdot \text{LOG}(1+1/d)$  percent. What if we instead sum for each bin the values of the numbers falling within and compare those nine sums? Such a process is called Summation Test, as it divides all amounts within the data set into digital compartments; distributing them to the various relevant digits according to leadership. In other words, the test combines an analysis of digits with their associated amounts by summing all actual amounts (not occurrences of digits) along digital lines. For example, if done along first-order digital proportions then all actual numbers within the data beginning with digit 1 are summed separately, another summation is performed for all actual numbers beginning with digit 2, and so forth, yielding nine different sums. If done along first-two-digits combinations (FTD), then all numbers beginning with digit combination 10 are summed separately; then all those beginning with combination 11 are summed separately, and so forth, until the last sum is calculated for those beginning with 99, yielding 90 different sums.

As an illustrative example of the definition and calculations of sums along digital lines, let us sum actual amounts along first-order digital lines for data pertaining to carbon dioxide emissions by 216 sovereign states and territories in 2008 (with the 217th entry standing for global emission amount). Data is downloaded from [http://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_carbon\\_dioxide\\_emissions](http://en.wikipedia.org/wiki/List_of_countries_by_carbon_dioxide_emissions). The given

6,427,875		31,592,984		2,402,788		7,021,885		7,870,826		300,705		8,142,402		127,649		299,330	
11	12,776	22	308	4	59	62	70	814	92								
103	12,835	29	389	411	59	66	77	851	99								
125	14,052	202	396	414	524	609	704	891	913								
128	15,130	213	3,146	422	539	649	708	8,016	920								
128	17,099	238	3,150	425	557	682	733	8,328	9,076								
143	17,158	246	3,542	447	576	6,113	7,015	8,328	94,660								
158	18,291	249	3,748	495	594	6,208	7,107	8,592	95,756								
161	102,936	260	3,953	4,067	5,203	6,219	7,591	8,672	97,814								
176	104,880	282	3,968	4,117	5,302	6,465	71,598	83,157									
180	111,304	2,057	31,276	4,118	5,548	6,912	73,109										
191	116,996	2,109	31,419	4,331	50,539	62,816	76,743										
198	124,905	2,156	32,295	4,602	54,638	67,700	78,371										
1,093	127,384	2,200	33,095	4,774	56,310	67,726	786,660										
1,228	155,066	2,230	37,557	4,815	56,512	68,478	7,031,916										
1,254	163,178	2,288	37,664	4,840	58,331												
1,335	169,533	2,314	38,573	4,976	509,170												
1,335	173,750	2,435	316,066	40,392	522,856												
1,353	192,378	2,439	323,532	40,535	538,404												
1,393	1,208,163	2,472	329,286	43,604	544,091												
1,419	1,708,653	2,560	376,986	45,749	5,461,014												
1,525	1,742,698	2,816	393,220	46,025													
1,533		2,816	399,219	46,527													
1,856		21,382		47,139													
1,889		21,617		47,840													
1,911		22,479		47,906													
1,918		23,304		49,050													
1,936		23,384		49,772													
1,951		24,371		49,920													
1,999		25,013		49,934													
10,392		26,826		406,029													
10,502		208,267		433,557													
10,594		210,321		435,878													
10,895		236,954		445,119													
11,764		258,599		475,834													
11,815		283,980		4,177,817													
11,914		285,733															
12,204		29,888,121															

Figure 3.8 Summation Test for Data on Carbon Dioxide Emissions Worldwide

value for each country represents annual emission in thousands of metric tons. First leading digits here are {26.7, 17.1, 10.1, 16.1, 9.2, 6.5, 6.5, 4.1, 3.7}. The table shown in Fig. 3.8 lists the entire data set of 217 values, displayed along first-order digital lines. Each number in Fig. 3.8 represents a single country (or rather its pollution level) plus a value for global emission. The amounts displayed on the top line are the nine sums allocated to each of the nine first-order digits. Of note here is that data in each of the nine compartments arises from approximately four to five different IPOT sub-intervals (i.e. each compartment has amounts with four to five different orders of magnitude). Clearly, the overall trend is towards decreasing sums for higher digits, albeit in an irregular manner. It is very hard to envision sums along digital lines for this data set as having some supposed equality.

Yet, these sums are mistakenly thought to be approximately equal for any honest data that is assumed to be Benford or nearly so. Thus the test is wrongly promoted in accounting and financial circles as an extra digital/quantitative test with the ability to flag any company providing fraudulent data by the simple lack of this

supposed equality. The fact that this ‘test’ is erroneous and misleading can be easily confirmed empirically, and results show that not a single company, honest or fraudulent, comes up with anywhere near sum-equality! Typically, sum for digit 1 is not double the sum for digit 9, nor tripled, but rather it’s always greater by a factor of at least four or five! In other words, sums are always high for the low digits, and are always low for the high digits, just as Benford’s Law itself dictates for occurrences of counts. The intuition and motivation though behind such sum-equality assertion can be clearly understood whenever data (having  $N$  elements) is artificially assumed to lie between adjacent IPOT values, such as **(10, 100)** for example. For data restricted to such an interval, there are quite numerous numbers to sum within (10, 20) where digit 1 dominates (30.1% to be exact), all having exceptionally low values (averaged 15). There are very few numbers to sum within (90, 100) where digit 9 dominates (4.6% to be exact), all having much higher values (averaged 95). Therefore perhaps there is a perfect trade-off and things might cancel out exactly. Sum along digit 1 is approximately  $(N*0.301)*15 = N*4.52$  while sum along digit 9 is approximately  $(N*0.046)*95 = N*4.37$ , hence the supposed equality. In other words, summing along digit 1 yields many low-quality numbers; while summing along digit 9 yields very few high-quality numbers, all of which suggests some kind of grand trade-off. For those extremely rare cases of logarithmic data sets and distributions confined exactly within such a ‘narrow’ range [i.e. between any two adjacent IPOT points such as (1, 10) for example], sums along digital lines are indeed equal, and this has been rigorously proven mathematically by Pieter Allaart.

This assertion can be demonstrated by hypothetical data set falling on (10, 100), such as: {10, 11, 12, 13, 15, 16, 18, 20, 23, 25, 28, 31, 33, 35, 43, 47, 53, 55, 61, 62, 79, 89, 97}. First digits are {30.4, 17.4, 13.0, 8.7, 8.7, 8.7, 4.3, 4.3, 4.3}, which is nicely near the logarithmic as much as can be for such small data size. Summing along first-digits lines we get:

$$10 + 11 + 12 + 13 + 15 + 16 + 18 = \mathbf{95}$$

$$20 + 23 + 25 + 28 = \mathbf{96}$$

$$31 + 33 + 35 = \mathbf{99}$$

$$43 + 47 = \mathbf{90}$$

$$53 + 55 = \mathbf{108}$$

$$61 + 62 = \mathbf{123}$$

$$79 = \mathbf{79}$$

$$89 = \mathbf{89}$$

$$97 = \mathbf{97}$$

Sums are nearly equal, but not sufficiently so since the data set is extremely small in addition to not being sufficiently close to the logarithmic. Much larger and nearly perfectly logarithmic data set on (10, 100) should show an almost exact sum-equality, consistent with Allaart's assertion. Yet surely real-life random data sets almost never fall within a narrow adjacent IPOT interval such as (1, 10), nor do they fall exactly between two non-adjacent IPOT points such as (1, 10000) in a deterministic exponential growth fashion with a constant logarithmic behavior throughout lacking digital development whatsoever. Rather, real-life data falls along multiple and partial IPOT sub-intervals, and always with a digital development pattern of increasing skewness, manifesting its digital configuration quite differently on the left, center, and right regions, resulting in inequality between sums on most of those regions. The aggregate sum along digital lines over all these regions combined is not equal, and no trade-off exists here whatsoever. Digital development implies that low digits are relatively weaker on the left where values are small and matter less, and that they are relatively much stronger on the right where values are much larger and matter much more, therefore the overall result is that low digits earn much higher overall sum than do high digits. In other words, since the region on the extreme far right of high values is the most crucial one for the overall sum, the fact that low digits are strongest there on the right implies that sums are not equal but are rather tilted in favor of the low digits. That perfect trade-off in digital development afforded to counts of numbers along digital proportions is not granted to amounts!

To demonstrate how digital development typically affects sums, let us consider (a more realistic) hypothetical data set falling on (10, 1000), with an approximate digital equality on (10, 100), and severe skewness on (100, 1000) in favor of low digits, such as:

{12, 14, 22, 27, 33, 36, 44, 49, 52, 55, 67, 68, 72, 77, 86, 88, 94, 97,  
122, 131, 137, 144, 162, 169, 183, 189, 193, 196, 198, 201, 234, 244,  
265, 269, 287, 302, 343, 387, 411, 456, 560, 623, 724}

First digits are {30.2, 18.6, 11.6, 9.3, 7.0, 7.0, 7.0, 4.7, 4.7}, which is nicely near the logarithmic as much as can be for such small data size. Summing along first-digit lines we get {1850, 1549, 1101, 960, 667, 758, 873, 174, 191}. Sum for digit 1 (i.e. 1850) is about 10 times the sum for digit 9 (i.e. 191)! Clearly sums here are not equal, but rather are highly skewed in favor of low digits, since low digits are the ones dominating the sub-interval (100, 1000) where it matters the

most as far as quantities are concerned. What happens on (10, 100) is by far less significant to sums of quantities.

Indeed **all** summation tests on actual statistical and random data relating to accounting and financial data, census data, single-issue physical data, and so forth, show a strong and consistent bias towards higher sums for low digits, typically by a factor of 5 to 12 approximately in the competition between digit 1 and digit 9. There is not a single exception!

In order to demonstrate the typical (uneven) behavior of these nine sums occurring in all random data, let us perform the summation test on the (almost perfectly) logarithmic U.S. Census data of Populations of Cities and Towns Incorporated. The largest population value of 8,391,881 belonging to New York City is considered an outlier, standing shoulder and above the rest of the other cities, and potentially swaying calculated sums a great deal. Hence NYC is being omitted altogether in this summation test. Figure 3.9 depicts the relevant chart showing clearly non-uniform sum series. In fact, this sum proportion does remind us a great deal of Benford's Law itself (except for the vertical scale)! Sums for digits 1 and 9 differ by a factor of 5.9, while in Benford's Law these proportions differ by a factor of 6.6, which is quite close. Clearly, summation test here did not

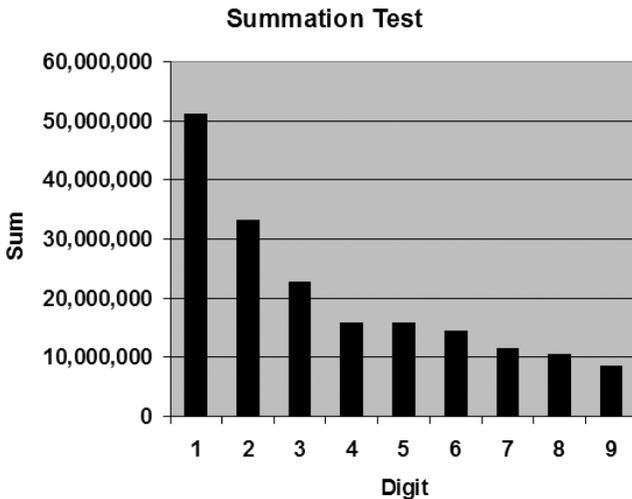


Figure 3.9 Sums Along First Digits, U.S. Populations Centers (New York City Excluded)

reveal any of that supposed equality of sums along digital lines. Rather, the opposite occurred, namely that low digits are not only blessed with higher existential proportions, but also live much richer lives and earning most of available amounts and resources.

Figure 3.10 depicts summation along the first-two digits for U.S. Cities and Towns Population data. The picture may initially appear a bit messy, yet an overall downhill trend is clearly and unmistakably visible, with sums differing roughly by a factor of six between those belonging to FTD near 99 and those belonging to FTD near 10.

Very similar results and charts relating to summation test are empirically obtained for numerous other real-life **random** data sets without a single exception, strongly refuting this erroneous dogma in equality of sums along digital lines. The only exception is found in **deterministic** data of the exponential growth series type, where sums along digital lines are indeed approximately equal, a fact which is consistent with Allaart's proof. Real-life data sets though are very rarely of the deterministic exponential growth series type.

A separate chapter with detailed analysis and theoretical discourse on summation test and its related issue of Sum-Invariant Characterization of the Law is given in the fourth section.

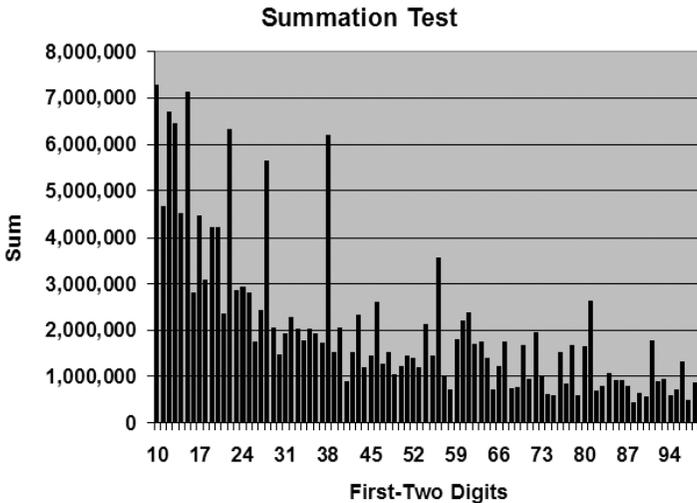


Figure 3.10 Sums Along First-Two Digits, U.S. Populations Centers (NYC Excluded)

## SUMMATION TEST IN THE CONTEXT OF FRAUD DETECTION

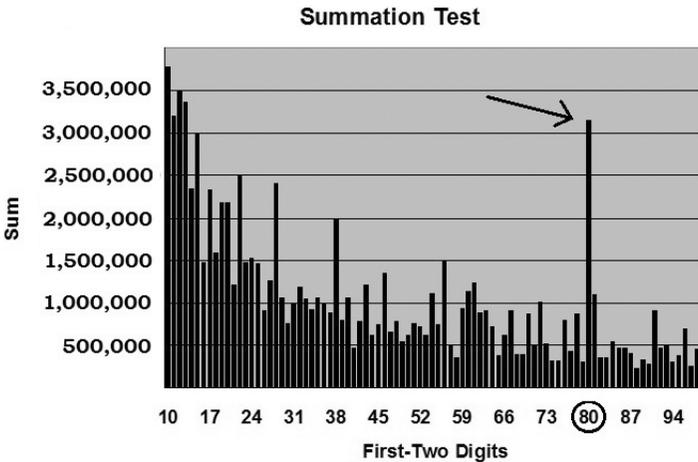
---

---

The knowledge that sums along digital lines in all random data sets are not equal but rather are heavily skewed in favor of low digits can actually be utilized in summation tests for forensic digital analysis in the context of fraud detection. The auditor should expect approximately a factor of 4 to 12 in the ratio of sum along digit 1 to the sum along digit 9 for the first order. Any unusual summation pattern very different from the expected one should trigger suspicion and further examination. Summation along first-two-digits (FTD) chart should also be included as it helps in pointing out more specifically to where unusually large sums had occurred to be further examined for possible fraud. The chart in Fig. 3.11 depicts the hypothetical outcome of summation test along FTD lines for expense account of a particular company. The overall downward trend seems quite authentic, but the large sum at 80 is suspicious and needs further investigation. Its magnitude is quite large relative to the much lower magnitudes in the neighborhood around 70–90. The magnitude seen here for the 80 combination would not raise an eyebrow had it occurred earlier on the left in the neighborhood around 10–15. Summation tests require a lot of subjective judgment (and experience) on the part of the forensic analyst; certainly there exist no critical cutoff points or confidence intervals to guide us.

It is instructive to note that the spike at 80 on the FTD summation chart could have its origin in a single large transaction of \$8,000,000, or equivalently from 100 transactions of \$80,000 each, and there is no way of knowing this from the outcome of the summation test alone. The auditor may use the value repetition test perhaps to verify which scenario is the real one. Digital FTD chart could also reveal whether or not digit combination 80 itself is overrepresented in the data (as counts of digits, not as sums of amounts), indicating perhaps that the 100 entries of \$80,000 scenario is the one that has actually occurred here. For very large account size and especially when the auditor wishes to pinpoint exactly where those unusual large sums occur, it is suggested to plot **sums along First-Three-Digits line!**

Whenever the digital analyst evaluating revenue or expense amounts encounters a nearly flat and uniform summation chart that gives roughly equal sums to all the relevant digits or digit combinations without any downward trend, this should raise strong suspicion that the account was concocted in a fraudulent way. Even more suspicion is raised when an upwards trend is detected. A rare exception to the above statements is when instead of revenues and expenses, the analyst is examining yearly balances of some trust fund dating back say to the French Revolution era, and sitting fixed at a certain secured British bank without any withdrawals or deposits, surviving world wars, revolutions, upheavals, and other calamities, calmly growing at a steady 5% annual interest rate for about two centuries. Sums along digital lines in the deterministic exponential growth series case are approximately equal, but since it takes large number of elements from any exponential growth series to manifest its logarithmic property, numerous years should be considered for such sum equality to hold approximately.



**Figure 3.11** Summation Test Applications in Forensic Analysis and Fraud Detection

## METHODS IN DIGITAL DEVELOPMENT PATTERN DETECTION

---

---

In this chapter an algorithm for the precise measurement and detection of digital development pattern in data sets shall be detailed. The motivations and explanations as to why only a partition of the entire range of data along IPOT sub-intervals is capable of showing the pattern shall be explored in the theoretical fourth section — Conceptual and Mathematical Foundations.

The following set of eight procedures allows us to decide on development.

- (1) A partition is constructed out of the entire range of data into the relevant sub-intervals bordered by adjacent integral powers of ten, such as  $[0.1, 1)$ ,  $[1, 10)$ ,  $[10, 100)$ ,  $[100, 1000)$ , and so forth. It is arbitrarily decided to have all left edges open and all right edges closed, as in  $[L, R)$ , so as not to overlap any points. As an example, suppose the minimum value in the data set is 3, and the maximum value is 7,468, then the leftmost sub-interval in the partition is  $[1, 10)$ , and the rightmost sub-interval is  $[1000, 10000)$ .
- (2) Once partition is made, first-digit distribution is calculated separately for each sub-interval, treating them as completely distinct data sets.
- (3) A summary table is constructed showing digital results from all sub-intervals, including the associated weight each sub-interval earns within the entire partition in terms of amount of data falling within it. Almost always, sub-intervals around the center contain most of the data, while very little data falls within the rightmost and the leftmost sub-intervals. The table shown in Figure 1.38 regarding digital development pattern for the 2012 earthquake data set may serve as an example of development summary table (except that IPOT data values such as 10, 100, 1000, and so forth, always belong to the sub-interval on their right to avoid overlapping as in our convention  $[L, R)$  of partitioning). Once the complete table of digital proportions on all sub-intervals is available, the digital analyst should be able to get an approximate idea about development by simply glancing at the table. Yet, it is necessary to

formalize and automate an algorithm in order to obtain an exact measure of development.

- (4) As a rule of thumb, any sub-interval on the extreme left or right edges with less than 0.1% of overall data (one tenth of one percent) should be designated as an outlier and be totally excluded from subsequent calculations in order to avoid distortions in overall result arising from very few erratic numbers. If the two leftmost or even the three leftmost sub-intervals (say) are both/all below that 0.1% threshold portion then both/all should be excluded from further analysis. Once these data-anemic and insignificant sub-intervals on the edges are eliminated, the plan is then to listen carefully to the digital messages coming from the rest of the (proper) sub-intervals in an equal measure, regardless of how much data they contain.
- (5) The definition of skewness over and above the Benford condition is given by the observed percents of digit 1 and digit 2 minus the Benford default sum of 47.7% for both these digits. In other words, the proportion of digits 1 and 2 in the observed data set for the sub-interval under consideration over and above their rightful legal allocation of 47.7% expected by the law. This measure is called '**Excess Sum Digits 1 & 2**' and abbreviated as **ES12**.

$$\text{ES12} = [\text{observed \% of digit 1} + \text{observed \% of digit 2}] - [\text{theoretical \% allocation for 1 and 2}]$$

$$\text{ES12} = [\text{observed \% of digit 1} + \text{observed \% of digit 2}] - [100 * \log(1+1/1) + 100 * \log(1+1/2)]$$

$$\text{ES12} = [\text{observed \% of digit 1} + \text{observed \% of digit 2}] - [47.7\%]$$

ES12 varies from +52.3% to -47.7%. A positive ES12 reading implies that digits 1 and 2 pooled more than their normal logarithmic share and thus digital configuration is skewed in favor of the low digits (i.e. 1 and 2) even more so than the skewness configuration of the Benford condition.

- (6) The series of ES12 numerical quantities is calculated and displayed for all the sub-intervals, in order to express the distinct digital skewness conditions throughout the entire range. The table in Figure 3.12 depicts the variation of ES12 along these IPOT sub-intervals for the population data of U.S. cities and towns (the two other measures of SPD and Saville's Slope are intended only for the more theoretical analysis performed at the end of this chapter and which can be ignored by accountants and auditors interested more in the

Left Point	1	10	100	1,000	10,000	100,000
Right Point	10	100	1,000	10,000	100,000	1,000,000
	===	===	=====	=====	=====	=====
Digit 1	14.8	5.3	19.1	37.3	46.0	62.9
Digit 2	7.4	8.1	17.4	19.7	20.2	17.6
Digit 3	3.7	7.0	13.6	11.6	10.9	6.0
Digit 4	7.4	9.2	11.5	8.6	6.4	4.1
Digit 5	7.4	11.5	9.9	6.3	5.8	3.0
Digit 6	14.8	13.9	8.8	5.3	3.8	3.0
Digit 7	7.4	13.9	7.6	4.3	2.8	1.5
Digit 8	14.8	17.0	6.1	4.0	2.3	1.1
Digit 9	22.2	14.1	6.0	2.9	1.7	0.7
ES12	-25%	-34%	-11%	9%	19%	33%
SPD	-7.3	-8.2	-2.0	1.8	3.5	5.7
Slope	-0.1	-0.4	0.5	1.3	1.7	2.3
% of Data	0.14%	5.5%	42.0%	37.3%	13.6%	1.4%

Figure 3.12 ES12, SPD, and Saville’s Slope for each IPOT Interval — U.S. Population

practical aspect of the algorithm). The trend of increasing skewness as focus moves to the right is clearly demonstrated here by the almost steady increase in the value of ES12. [Note: defined sub-ranges seem to overlap at 10, 100, 1000, etc., yet those very few cities or towns with exact IPOT values within the data set were placed always on the sub-intervals to their right as in the rule [L, R) for the edges mentioned in procedure #1.]

- (7) When ES12 is plotted (vertically) versus a simple counting index for the sub-intervals (horizontally), we expect to observe a rising trend — reflecting the increase in skewness when focus shifts to the right side of the range of data. In order to carefully measure the magnitude of the rise (i.e. its steepness — which signifies digital development), Simple Linear Regression analysis is used. In conducting regression analysis, ES12 is to be considered as the dependent variable. An associated arbitrarily defined independent variable  $N_i$  — the natural numbers — is to serve as an index representing the sub-intervals, so that the first sub-interval, the leftmost one (if its data portion is at least 0.1%) is given the index value 1, the next sub-interval to the right of it is given the value 2, and so forth, until we reach the last index value called L for the rightmost sub-interval. L is simply the total number of relevant sub-intervals in the analysis for the particular data set in question, outliers excluded. For the U.S. population centers data set L is 6.

The Simple Linear Regression model is then:  $ES12_i = \mathbf{b} * N_i + \mathbf{a}$ , where  $\mathbf{b}$  is the slope and  $\mathbf{a}$  is the intercept. Figure 3.13 depicts the regression plot for the

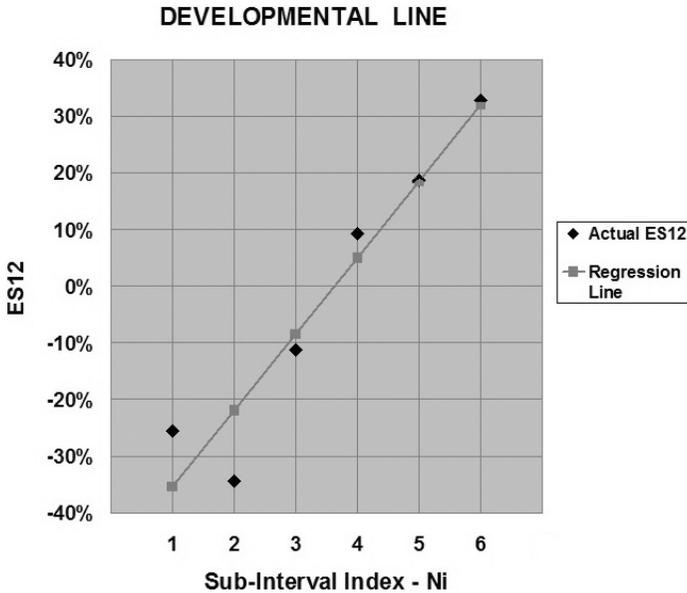


Figure 3.13 ES12 Developmental Line of the 1st Order Digits for U.S. Population Data

U.S. Population Census data on all cities and towns by applying ES12 measure of digital skewness. The points shown in Figure 3.13 are derived from the 6 values in Figure 3.12 in the row for ES12.

Simple Linear Regression of  $\{-25.5\%, -34.4\%, -11.2\%, 9.3\%, 18.6\%, 32.8\%\}$  on  $\{1, 2, 3, 4, 5, 6\}$  yields the slope  $\mathbf{b} = +13.5\%$  and the intercept  $\mathbf{a} = -48.8\%$ .

- (8) The regression line (to be called ‘**Developmental Line**’) fitting these points should have a decisive and significant positive value for the slope. Yet, our developmental algorithm does not attempt to directly confirm that the value of the slope is positive, it examines instead two related values which are considered more stable, consistent, and reliable. It examines whether the leftmost point on the regression line is sufficiently below 0, as well as that the rightmost point is sufficiently above 0. Absent any statistical theory for significance of the parameters of developmental line, no objective critical values exist to enable us to set firm cutoff points to decide whether or not there is a significant increase in skewness across sub-intervals (namely development). Yet, the forensic data analyst could still utilize empirical threshold values based on the studies and careful observations of actual results from a diverse set of real-life random data performed by the author [*all showing a definite development without*

a single exception]. Such accumulated knowledge about development enables us to flag any accounting or financial data as suspicious whenever its developmental line shows strong deviation from the norm.

Two extreme values on the regression line (the two endpoints) are calculated by letting independent index variable  $N_i$  take the minimum value of 1 and the maximum value of  $L$ :

$$ES12(\text{left}) = b*1 + a$$

$$ES12(\text{right}) = b*L + a.$$

Reported data passes developmental test if the following two conditions are true:  
 $ES12(\text{left}) < -25\%$  and  $ES12(\text{right}) > +25\%$

The constraint of this test constitutes the EMPIRICAL LAW OF DEVELOPMENT, true across all random and statistical data types, logarithmic as well as non-logarithmic ones. Failure to pass the test strongly hints at the possibility of fraudulent activity in data reporting.

For U.S. Population Centers data set, the vertical ES12 coordinates of the two endpoints of Developmental Line can be clearly ascertained visually from Figure 3.13. Calculating exactly these two values, we get:

$$ES12(\text{left}) = b*1 + a = +13.5\%*1 + -48.8\% = -35.3\%.$$

$$ES12(\text{right}) = b*L + a = +13.5\%*6 + -48.8\% = +32.2\%.$$

$$ES12(\text{left}) = -35.3\% < -25\% \quad \text{and} \quad ES12(\text{right}) = +32.2\% > +25\%$$

Therefore U.S. Population data set easily passes developmental test.

**This is for all practical purposes the END OF THIS CHAPTER for accountants and auditors interested in applying this forensic digital algorithm in fraud detection.**

**The REST OF THIS CHAPTER is a more complex treatment of the theoretical development of the method itself as well as three additional digital developmental tests.**

- (4) As a rule of thumb, any sub-interval on the edges with less than 0.1% of overall data should be designated as an outlier and be totally excluded from subsequent analysis. In addition, the rest of the sub-intervals are to be considered in equal measure, regardless of their data-portion weights. These two rules should be strictly followed, and in spite of potentially strong objections and counter arguments that such complete elimination of digital configurations on the left and right edges in development analysis due to severe scarcity of data can be avoided by

simply applying data-portion weights to each sub-interval, thus avoiding any undue influence in results from data-anemic sub-intervals. Yet, it is strongly suggested here **not** to apply data-portion weights to individual sub-intervals in all subsequent techniques in digital development analysis. There is a need to obtain the direct or raw (unweighted) leading digits configurations from each sub-interval, in order to have them contribute equally to the overall measure of development, regardless of how much data falls within each one. This is so since development often manifests itself almost from the very left and from the very right edges on the margin of data, hinting at the correct digital development form. Nonetheless, it is not wise to include highly data-deficient sub-intervals and let very few points on the left or right edges dramatically sway result; hence the 0.1% outlier rule of elimination for the two extreme edges. The arbitrary cutoff point of 0.1% is a reasonable compromise between these two opposing goals. Put another way, only such measuring approach leads to (nearly) steady and consistent empirical observations of digital development across multiple real-life data sets. Had we allowed in any sub-interval on the left or right, even with data less than 0.1% portion, and/or had we applied weights in the analysis, no universal measure of digital development (of almost exact numerical value) could have been observed and stated.

- (5) Once the complete table of digital proportions on all proper sub-intervals is available, the algorithm to construct a singular measure of development needs to be formalized and automated. To accomplish that, it is necessary first to create an exact numerical measure of digital skewness over and above the Benford condition so as to be able to assign each sub-interval such a value. To accomplish that in turn, it is necessary to decide upon which digits are to be designated 'high' and which are to be designated 'low', in order to give meaning to a statement such as 'digital configuration is skewed in favor of low digits'. Surely 1 and 2 are to be considered low digits, and surely 8 and 9 are to be considered high digits, but how do we classify digits 3, 4, 5, 6, and 7? Would any classification of digits here be considered arbitrary? The question is not merely linguistic or philosophical, but actual as it is relevant here in the definition and calculation of skewness. A naïve approach would be to attempt to equally divide all 9 digits relevant to the first order, so that 1, 2, 3, 4 would be considered low, and 6, 7, 8, 9 would be considered high; but then what should be done with the middle digit 5? An odd set of 9 cannot be divided into two equal subsets.

One simple way of resolving this dilemma is found by noting that according to Benford's Law itself in extreme generality "*digits are skewed in favor of the low*

ones”, with digits 1, 2, and 3 obtaining higher leadership than what would have been obtained under the naïve assumption of 11.1% equality for all 9 digits. Therefore, even digit 3 with its allocated 12.5% according to the law is just a bit higher than that supposed 11.1% of equality, and so should also be considered a low digit. The rest of the digits, 4 to 9, are to be considered high ones since they are all allocated less than that supposed 11.1% equality.

The third and by far the most attractive approach here is to designate **1 and 2 as low digits**, and **3, 4, 5, 6, 7, 8, 9 as high digits**, being the best partitioning of digits into two opposing camps in a probabilistic sense as per Benford’s Law itself. Since about half of real-life numbers are being led by digits 1 and 2 (30.1% + 17.6% = 47.7%), and roughly the other half are being led by 3, 4, 5, 6, 7, 8, 9 (12.5% + 9.7% + 7.9% + 6.7% + 5.8% + 5.1% + 4.6% = 52.3%), such classification is more natural to consider and it is facilitated via the law itself. One might claim that pitting digits 1 and 2 versus digits 3, 4, 5, 6, 7, 8, 9 would constitute a ‘fair’ tug of war, probability-wise. We shall adopt this last approach and consider {1, 2} the set of low digits.

Hence a reasonable definition of skewness over and above the Benford condition is given by the observed percents of digits 1 and 2 in the data set minus the Benford default sum of 47.7% for both digits, namely:

$$\text{ES12} = [\text{observed \% of digit 1 and 2}] - [\text{theoretical \% allocation for 1 and 2}].$$

An alternative measure of skewness, to be called “**Sum Percent Deviations**” and abbreviated as **SPD** is defined as the sum of  $(O_i - B_i)/B_i$  for digits 1 and 2, and  $(B_i - O_i)/B_i$  for digits 3, 4, 5, 6, 7, 8, and 9, where  $O_i$  denotes the Observed proportion of numbers with first digit  $i$  within the particular sub-interval in question, and  $B_i$  denotes the proportion of digit  $i$  according to Benford’s Law, that is:

$$\begin{aligned} \text{SPD} = \text{Sum Percent Deviations} &= (O_1 - 0.301)/0.301 + (O_2 - 0.176)/0.176 \\ &+ (0.125 - O_3)/0.125 + (0.097 - O_4)/0.097 + (0.079 - O_5)/0.079 \\ &+ (0.067 - O_6)/0.067 + (0.058 - O_7)/0.058 + (0.051 - O_8)/0.051 \\ &+ (0.046 - O_9)/0.046 \end{aligned}$$

It is noted that for a perfectly Benford data set SPD is zero. A positive SPD value indicates that data is more skewed in favor of the low digits than as in the Benford condition. A negative SPD value implies in general one of the following three scenarios: (A) Digits are not as skewed in favor of the low digits as in the Benford condition, but still low digits are favored. (B) There is a near digital equality. (C) High digits win slightly or strongly.

For example, for the distribution severely skewed in favor of low digits such as {45%, 22%, 15%, 6%, 3%, 3%, 3%, 2%, 1%}

$$\begin{aligned} \text{SPD} &= (0.45 - 0.301)/0.301 + (0.22 - 0.176)/0.176 + (0.125 - 0.15)/0.125 \\ &+ (0.097 - 0.06)/0.097 + (0.079 - 0.03)/0.079 + (0.067 - 0.03)/0.067 \\ &+ (0.058 - 0.03)/0.058 + (0.051 - 0.02)/0.051 + (0.046 - 0.01)/0.046 \\ \text{SPD} &= 0.495 + 0.250 + -0.200 + 0.381 + 0.620 + 0.552 + 0.483 + 0.608 + 0.783 \\ \text{SPD} &= +3.97 \end{aligned}$$

$$\begin{aligned} \text{ES12} &= [\text{observed \% of digit 1} + \text{observed \% of digit 2}] - [47.7\%] \\ \text{ES12} &= [45\% + 22\%] - [47.7\%] = +19.3\% \end{aligned}$$

For a distribution with nearly digital equality such as {13%, 13%, 11%, 11%, 11%, 15%, 8%, 9%, 9%}

$$\begin{aligned} \text{SPD} &= (0.13 - 0.301)/0.301 + (0.13 - 0.176)/0.176 + (0.125 - 0.11)/0.125 \\ &+ (0.097 - 0.11)/0.097 + (0.079 - 0.11)/0.079 + (0.067 - 0.15)/0.067 \\ &+ (0.058 - 0.08)/0.058 + (0.051 - 0.09)/0.051 + (0.046 - 0.09)/0.046 \\ \text{SPD} &= -0.568 - 0.261 + 0.120 - 0.134 - 0.392 - 1.239 - 0.379 - 0.765 - 0.957 \\ \text{SPD} &= -4.58 \end{aligned}$$

$$\begin{aligned} \text{ES12} &= [\text{observed \% of digit 1} + \text{observed \% of digit 2}] - [47.7\%] \\ \text{ES12} &= [13\% + 13\%] - [47.7\%] = -21.7\% \end{aligned}$$

One minor drawback in the definition of SPD is that the last seven terms are somewhat determined by the first two, and vice versa. If the proportion of digit 1 is 40% and that of digit 2 is 25%, then exactly 35% is allocated to digits 3 to 9, and most or all of the last seven terms of SPD are likely to come out positive, adding a bit more to the positive terms of  $(0.40 - 0.301)/0.301 + (0.25 - 0.176)/0.176$  in the overall sum and reinforcing the positive result. In other words, the first two terms and the last seven terms are not independent. Nonetheless, there is still a variety of ways to allocate percents to those remaining seven terms. Moreover, the overall ability of SPD definition to express skewness is not in doubt.

It is noted that SPD of the entire data set itself always equals the sum of SPDs over all sub-intervals weighted by the proportion of overall data within each one. In other words, SPD of the data set itself is the weighted average of the various SPDs of all the sub-intervals (including those with less than 0.1% of data). No matter how we partition the range, as long as we do not omit or overlap any segment, the weighted sum of all SPDs is a constant, partition-invariant, and equals to SPD of the entire data set.

The definitions of ES12 and SPD should not be considered arbitrary. Surely there are many other reasonable ways to measure and express the very same concept of digital skewness over and above the Benford condition. It could have been defined for example as  $\{1, 2\}$  versus  $\{8, 9\}$  utilizing just the four extreme digits, such as in the similar expression  $(O1 - 0.301)/0.301 + (O2 - 0.176)/0.176 + (0.051 - O8)/0.051 + (0.046 - O9)/0.046$  for SPD. We could also pit  $\{1, 2\}$  against  $\{6, 7, 8, 9\}$ , or we may wish to pit  $\{1, 2, 3\}$  against  $\{7, 8, 9\}$  for example. Yet all other such alternative definitions if constructed reasonably enough to express the concept of skewness over the Benford condition yield the same conclusion regarding digital development for any real-life random data set [actual empirical experimentations by the author strongly confirmed this fact for several alternative definitions of skewness]. This is so since real-life data sets always show two very distinct development styles, being either random with that ubiquitous digital development, or deterministic exponential-growth-like having a very steady logarithmic behavior throughout its entire range. Hence, including or excluding a middle digit or two, or utilizing other (similarly constructed) algebraic expressions does not change the overall conclusion in the least, even though exact output values surely come out differently (while their interpretations lead to identical conclusions).

A superior and by far the most elegant measure of skewness over and above the Benford condition can be found in the value of the slope in Saville's regression scheme, which designates data as 'more skewed than Benford' whenever it is larger than 1, and 'less skewed than Benford' whenever it is less than 1. Saville's scheme allows us not to have to decide which digits are termed low and which digits are termed high. Digital development in random data implies that Saville's Slope on those mini sub-intervals between adjacent IPOT points experience the transition from being either negative or near zero on the left, to about 1 around the center, then rising decisively above 1 on the far right part of the range.

(6) A single numerical quantity is calculated and displayed for each IPOT sub-interval measuring digital skewness condition by calculating any one of the following three items:

(I) Excess Sum Digits 1 and 2 (ES12), (II) Sum Percent Deviations (SPD), or (III) Saville's Slope.

(7) In conducting regression analysis, ES12, SPD, or Saville's Slope values are to be considered as the dependent variable. Our Simple Linear Regression model for SPD is  $SPDi = \mathbf{b} * Ni + \mathbf{a}$ , where  $\mathbf{b}$  is the slope and  $\mathbf{a}$  is the intercept. Similar models are defined for ES12 as well as for Saville's Slope.

- (8) The regression line fitting these points should have a positive slope, given that the data is random in nature as opposed to the deterministic. We shall call any of the above three possible lines ‘Developmental Line’ (of the ES12, SPD, or Saville’s Slope variety).

It would not be correct to judge compliance or non-compliance with development based solely on the values of ES12, SPD, or Saville’s Slope on the rightmost sub-interval and on the leftmost sub-interval, while ignoring the central region. Rather, Simple Linear Regression analysis is needed in aiding such decisions. It is necessary to use regression so as to smooth out any abrupt and local deviation on the far right or on the far left, and to take the overall developmental pulse coming out from **all** IPOT sub-intervals.

It is not possible to simply take the value of the slope of developmental line as the measure of development. This is so because empirical studies show that the slope varies wildly between data sets, depending on the number of relevant IPOT sub-intervals; namely that it depends on the value of L. For a typical random data set with large L, the slope is small and mild since overall variation in SPD, ES12, or Saville’s Slope is divided and shared by many sub-intervals, gradually increasing in many small steps. For a data set with very small L value such as 2 or 3, the slope is relatively large, since overall variation in SPD or ES12 is short and dramatic, accomplished in just two or three steps. Hence the slope itself cannot be utilized as a reliable and consistent (universal) measure of development.

**Empirically, what is approximately invariant across almost all random data sets is the universal positive level of SPD/ES12/Slope on the rightmost point of the regression line (Developmental Line), and its universal negative level on the leftmost point.** The gap (difference) between these two values is not quite steady, but there is always a minimum rise in SPD above +5 on the right, and always a minimum fall below -4 on the left. Alternatively, there is always a minimum rise in ES12 above +25% on the right, and always a minimum fall below -25% on the left. These two extreme values on the regression line are calculated by letting independent index variable  $N_i$  take the values 1 and L:  $SPD/ES12(\text{left}) = b*1 + a$ , and  $SPD/ES12(\text{right}) = b*L + a$ .

The digital analyst should insist on finding out that SPD(left) or ES12(left) is sufficiently small, as well as that SPD(right) or ES12(right) is sufficiently large. It is recommended that four tests are to be performed in order to identify reported financial and accounting data as either honest and in conformity with its supposed random and statistical developmental nature, or as possibly fraudulent. Failure to

pass any one of the four tests should identify the data as suspicious and due further investigation. It does not matter which measure of skewness is selected, ES12 or SPD, but each measure has different empirical intervals of acceptance. Empirical studies on Saville's Slope behavior have not been performed yet, thus this measure shall be omitted in the following discussion.

### Test I:

Reported data passes Test I if the following two conditions are true:

$$\begin{aligned} \text{ES12}(\text{left}) < -25\% & \quad \text{and} \quad \text{ES12}(\text{right}) > +25\% \\ \text{SPD}(\text{left}) < -4 & \quad \text{and} \quad \text{SPD}(\text{right}) > +5 \end{aligned}$$

The above set of numerical constraints (Test I) constitutes the first part of the empirical law of development, true across all random and statistical data types, logarithmic as well as non-logarithmic.

(9) Yet, the neurotic data analyst wishes to take one further step here to ensure that he or she has not been blinded by the seductive power of regression theory, recalling the many pitfalls associated with the method of linear regression. It is necessary to pay attention to the possibility of false negative occurrences, where the central region (bulk of data) is roughly skewed-neutral and nearly always logarithmic (contrary to developmental theory), and where only the two extreme sub-intervals on the right and on the left are responsible for any perceived development, thus managing to manipulate regression line to appear as if the data in its entirety is complying with the supposed pattern. In other words, it is necessary to confirm that the line is climbing upwards throughout (as is the case with almost all honest random data), and that it is not climbing only due to the edges. To ensure that development really exists throughout, we sum up the absolute values of ES12 or SPD and insist that this sum be larger than some empirical cutoff value. If there is really no development and the logarithmic property is approximately true everywhere around the center, then such sum should be quite low since ES12 and SPD values tend to be near 0 on most sub-intervals.

### Test II:

Reported data passes Test II if the following is true:

$$\text{Sum of absolute values of all ES12} > 110\%$$

$$\text{Sum of absolute values of all SPD} > 18$$

The above set of numerical constraints (Test II) constitutes the second part of the empirical law of development, true across all random and statistical data types, logarithmic as well as non-logarithmic.

- (10) The above discussion about digital development focused exclusively on the first-order leading digits, although all this can be considered and performed on the second-order leading digits as well with only few modifications. Yet for the second order development analysis, quite often the random element overwhelms the systematic one, obscuring results, unless data set is sufficiently large so that clarity in the pattern of digital development can be observed. This is so because of the more delicate differences within the second-order digits, being that they are not nearly as skewed in favor of low digits as is the case for the first leading digits. For this reason, second-order digital development analysis should not be performed on small data sets.

Whenever second-order digital development analysis is performed, it is suggested to stick to a partition along adjacent IPOT as well, just as is done for the first-order test. This is so in spite of the fact that the full cycle of the second leading digits is much shorter, being merely 1-unit long, equally and consistently everywhere on the x-axis. For example, from 7.0 to 8.0 all second order digits possibilities get a chance to manifest themselves, with digit 0 second leading on [7.0, 7.1), digit 1 second leading on [7.1, 7.2), digit 2 second leading on [7.2, 7.3), and so forth. The reason for suggesting to test second order along first order lines of 1, 10, 100, and so forth, is due to the interdependencies of the orders, being that second order probabilities depend on first order probabilities; that second digits proportions are more skewed in favor of second-order low digits whenever first-order digits are also low. By selecting IPOT sub-intervals in which all first-order digits are equally represented such as [1, 10) and [10, 100), by extension we allow all second order digits equal opportunity to fully express themselves and their logarithmic property as well.

- (11) Which second-order digits should be designated as low and which as high, now that digit 0 is also included and all 10 digits are to be considered? Interestingly, all three considerations and arguments mentioned earlier for the designation of digits in the case of the first order, unite here in agreeing that **{0, 1, 2, 3, 4} should be considered low second digits, and {5, 6, 7, 8, 9} high ones**. A simple division of equal partition is surely such. Also,

Benford's second order probabilities for digits 0, 1, 2, 3, 4 are all above that 10% [naïve] equality in distribution, and probabilities for digits 5, 6, 7, 8, 9 are all below that 10% supposed equality. The sum of probabilities for 0, 1, 2, 3, 4 is 54.7%, while the sum of probabilities for 5, 6, 7, 8, 9 is 45.3%, and no other partition can yield more balanced or equitable totals closer to the ideal 50% proportions.

- (12) Hence a reasonable definition of skewness over and above the second order Benford condition on any given IPOT sub-interval is the observed percent of numbers being led by digits  $\{0, 1, 2, 3, 4\}$  minus the Benford default sum of 54.7% for all these 5 second order digits.

$$\mathbf{ES04} = [\text{observed \% of all second order digits } \{0, 1, 2, 3, 4\}] - [54.7\%]$$

This second order skewness measure, to be called '**Excess Sum Digits 0 to 4**' and abbreviated as **ES04**, varies from +45.3% to -54.7%.

An alternative measure would define '**Sum Percent Second Order Deviations**' and abbreviated as **SPD2** as the sum of  $(O_i - B_i)/B_i$  for digits 0 through 4, and  $(B_i - O_i)/B_i$  for digits 5 through 9, where  $O_i$  denotes the actual observed proportion of numbers with second digit  $i$  within the specific sub-interval in question, and  $B_i$  denotes the proportion of digit  $i$  according to the second order Benford's Law. That is:

$$\begin{aligned} \mathbf{SPD2} = \text{Sum Percent Second Order Deviations} \equiv & (O_0 - 0.120)/0.120 \\ & + (O_1 - 0.114)/0.114 + (O_2 - 0.109)/0.109 + (O_3 - 0.104)/0.104 \\ & + (O_4 - 0.1003)/0.1003 + (0.097 - O_5)/0.097 + (0.093 - O_6)/0.093 \\ & + (0.09 - O_7)/0.09 + (0.088 - O_8)/0.088 + (0.085 - O_9)/0.085 \end{aligned}$$

### Test III:

Reported data passes Test III if the following two conditions are true:

$$\mathbf{ES04}(\text{left}) < -2\% \quad \text{and} \quad \mathbf{ES04}(\text{right}) > +6\%$$

$$\mathbf{SPD2}(\text{left}) < -1.0 \quad \text{and} \quad \mathbf{SPD2}(\text{right}) > +1.0$$

### Test IV:

Reported data passes Test IV if the following is true:

$$\text{Sum of absolute values of all } \mathbf{ES04} > 15\%$$

$$\text{Sum of absolute values of all } \mathbf{SPD2} > 3$$

The table in Figure 3.14 depicts mini second-order digit distributions on IPOT sub-intervals for the census data on U.S. population centers. The leftmost

sub-interval on (1, 10) with only 0.14% of overall data was omitted here so as to focus solely on the sections where development is clear enough — since second order has a more subtle distribution. Certainly a clear second order development pattern is seen here in the direction of increasing skewness as focus moves to the right, just as was seen in the first order case. Simple Linear Regression of  $\{-3.6\%, -2.7\%, 2.8\%, 5.1\%, 12.7\%\}$  on  $\{1, 2, 3, 4, 5\}$  yields the slope  $\mathbf{b} = +4.0\%$  and the intercept  $\mathbf{a} = -9.3\%$  for ES04 skewness measure.

The regression plot in Figure 3.15 depicts ES04 Developmental Line for the data on U.S. population centers which easily passes both tests (III and IV) of the second order development, as follows:

$$ES04(\text{left}) = b*1 + a = +4.0\%*1 + -9.3\% = -5.3\%.$$

$$ES04(\text{right}) = b*L + a = +4.0\%*5 + -9.3\% = +10.7\%.$$

$$ES04(\text{left}) = -5.3\% < -2\% \quad \text{and} \quad ES04(\text{right}) = +10.7\% > +6\%$$

Therefore U.S. Population data set easily passes Developmental Test III.

Sum of absolute values of all ES04

$$= |-3.6\%| + |-2.7\%| + |2.8\%| + |5.1\%| + |12.7\%| = 26.9\% > 15\%$$

Therefore U.S. Population data set easily passes Developmental Test IV.

Forensic examination of a large variety of real-life random data regarding financial and accounting amounts, census data, single-issue physical phenomenon,

Left Point	10	100	1,000	10,000	100,000
Right Point	100	1,000	10,000	100,000	1,000,000
	====	=====	=====	=====	=====
Digit 0	9.7	9.9	13.0	13.8	24.0
Digit 1	8.7	10.6	12.3	12.1	11.6
Digit 2	10.7	10.7	11.3	13.0	14.2
Digit 3	10.1	10.5	10.7	10.6	8.6
Digit 4	11.8	10.1	10.1	10.4	9.0
Digit 5	10.1	9.6	9.4	8.7	7.9
Digit 6	10.1	10.0	9.2	9.6	5.2
Digit 7	9.9	9.3	8.1	8.6	7.1
Digit 8	9.7	9.8	8.4	6.9	6.0
Digit 9	9.1	9.3	7.4	6.5	6.4
	-----	-----	-----	-----	-----
ES04	-4%	-3%	3%	5%	13%
SPD2	-0.7	-0.5	0.6	1.0	2.5
% of Data	5.5%	42.0%	37.3%	13.6%	1.4%

Figure 3.14 ES04 and SPD2 of 2nd Order Development — U.S. Population Data

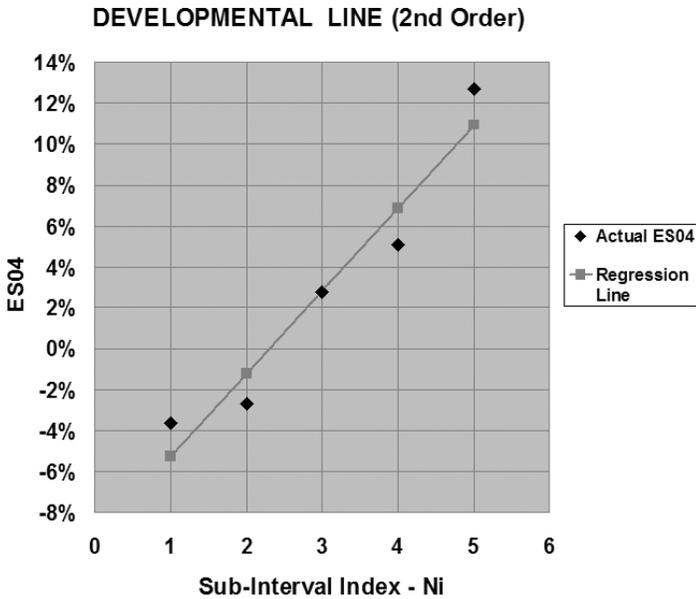


Figure 3.15 ES04 Developmental Line of the 2nd Order for U.S. Population Data

and many other data types, easily pass all four forensic development tests above. The cutoff points in all four tests were intentionally made in a liberal fashion so as to guarantee passage and compliance in all honest real-life random data sets.

Also of note here is that the left portion of digital equality and the right portion of extreme digital inequality typically represent a small portion of overall data, while the logarithmic-like center contains the bulk of the data. If a corporation files its income tax by providing data that not only is nicely logarithmic, but also on the face of it comes with the basic correct developmental style outlined here, but where most of the data lies on the two regions of digital equality and severe digital inequality, with a small portion residing in the logarithmic-like center, then suspicion could and should arise.

## CASE STUDY VIII: PRICE LIST OF A LARGE MANUFACTURER

---

Canford Audio (CA) PLC in the U.K. manufactures and retails electronic items. It specializes across the whole spectrum of audio, video, and data hardware and infrastructure products. Its website is <http://www.canford.co.uk/>. The icon “Download pricelist” at the top right area is selected with a choice of U.S. Dollar and .xls as the file format in MS-Excel. Prices are quoted on column H. Out of a total of 15,494 rows, 296 are designated as “POA” and are deleted along with four entries having the value of 0, leaving 15,194 entries for proper analysis.

Surely there can be no issue with honesty or fraud for this data set, since this is authentically the price list itself, as opposed to reported revenue data. Yet the price list shall be treated here as if it was revenue data in order to demonstrate fraud detection techniques. The enormous size of this price list renders it sufficiently large in a statistical sense to serve as an excellent case study. Analyses of the following five basic digital distributions shall be performed:

The 1st order digit distribution is:

CA PLC List — {28.8, 17.7, 14.2, 9.2, 8.1, 7.0, 5.3, 5.1, 4.6}

BL 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

CA PLC List — {11.1, 12.0, 10.9, 10.5, 10.4, 8.6, 10.6, 8.8, 7.7, 9.5}

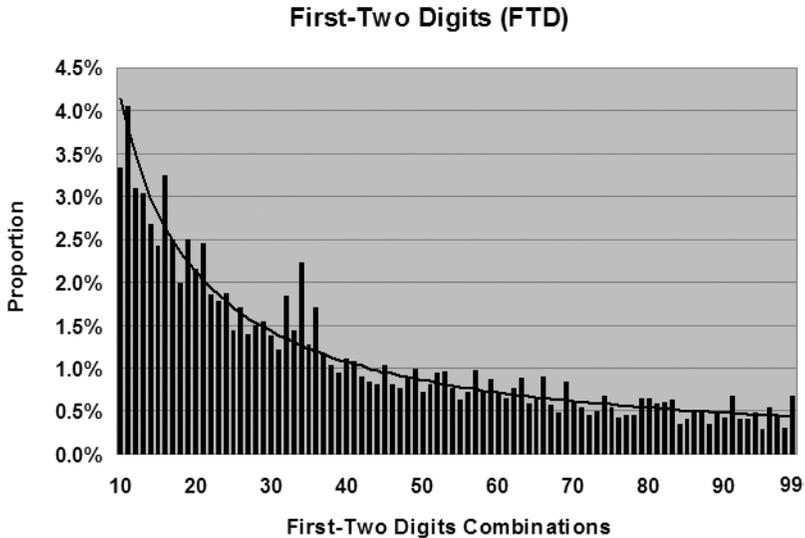
BL 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

The 3rd order digit distribution is:

CA PLC Listx — {10.92, 9.19, 10.45, 10.14, 9.36, 10.14, 10.16, 9.91, 10.75, 8.97}

BL 3rd Digits — {10.18, 10.14, 10.10, 10.06, 10.02, 9.98, 9.94, 9.90, 9.86, 9.83}

FTD chart depicted in Fig. 3.16 appears nicely logarithmic. Thirteen values between \$0.02 [two cents] and \$0.08 [eight cents] were omitted since they are too short digit-wise, having no second-order digit to consider. In the same vein, for the third order calculations, it is necessary to eliminate 584 values below



**Figure 3.16** First-Two Digits — Canford Audio PLC Price List

\$1.00 having no meaningful third significant digit, such as \$0.02, \$0.10, \$0.37, \$0.81, or \$0.99. LTD chart depicted in Fig. 3.17 on the other hand appears quite problematic, having six noticeable spikes above the 2% line. There are 3494 values between \$0.02 and \$9.98 which are too short digit-wise, and therefore must be omitted in the preparation of LTD chart. For example, \$9.98 should not be allowed to contribute 98 to the LTD chart, since 98 does not come with the digital equality of 1%; rather, 9 belongs to the second order and 8 belongs to the third order. Hence only 11,700 values from \$10.00 upwards were included in the LTD chart [treating the third order as digital equality in the approximate].

SSD of 1st digits yields the very low value of 5.3

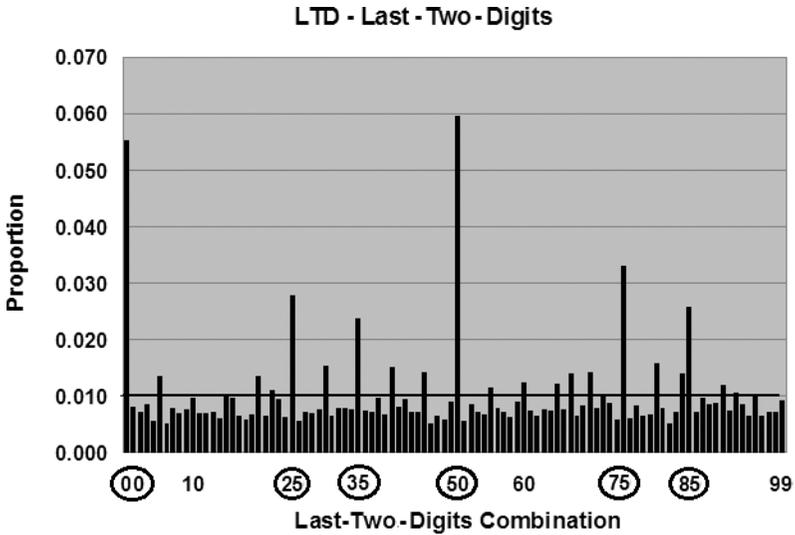
SSD of 2nd digits yields the very low value of 6.3

SSD of 3rd digits yields the very low value of 3.6

SSD of FTD yields the very low value of 4.4

SSD of LTD yields the relatively high value of 66.3

Saville Test for the first digits yields a slope of 0.96 and an intercept of 0.0043, indicating that the data set is perhaps very close to Benford in the first digits sense. This result is compatible with the low SSD value gotten for the first order.



**Figure 3.17** Last-Two Digits — Canford Audio PLC Price List

The Canford Audio PLC price list should not be considered non-logarithmic in spite of the considerable deviation from the law found in its LTD distribution. The reason is that in the special case of price lists, which are determined and created (intentionally) by the company's employees to some extent, there are valid exceptions to Benford's Law. As an example, let us consider one item in the price list of Canford, named 'NTI Acoustic Test Kit Pro Factory Recalibration' having the price tag of \$653.40. Surely the employee deciding on this exact price must consider the cost of the item to the firm, competitors' prices, and the best way to attract potential customers. He or she has very little leeway — if any at all — in the first digit which probably must be 6 regardless, expressing \$600. The first digit of 6 is strongly determined by forces outside the employee's control [reflecting cost on the downside as the item can't be sold for less than \$600 lest loss is incurred, and limits due to competition on the upside as the item can't be sold for more than \$700 lest customers are lost to competitors]. Yet the last two digits of 40 cents can surely be changed regardless of cost or even competition, hence there exists partial human-intentional factor in pricing. In fact, this is corroborated by the fact that all the spikes on the LTD chart are positioned at some very popular quotes for prices, such as 00, 25, 35, 50, 75, and 85. Moreover, the fact that the two truly biggest spikes of 00 and 50 are extremely popular in price quoting,

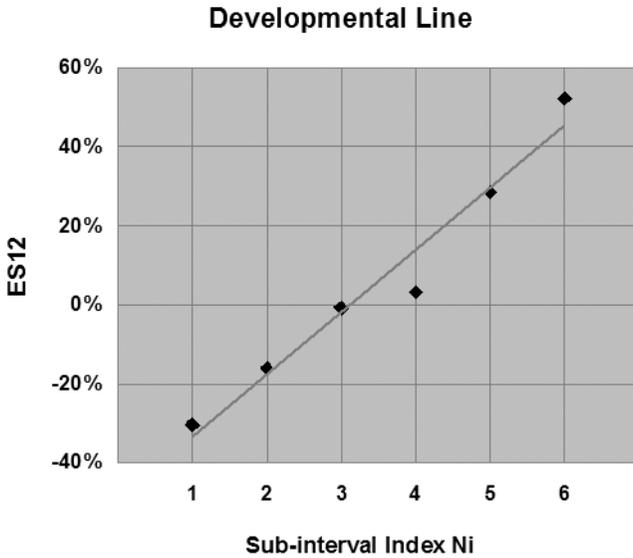
strongly corroborates the discussion above, and excuses Canford Audio for having significant deviation in its LTD chart.

Just in case the data happened to be provided by a scheming employee already aware of Benford’s Law (including all higher-order digital proportions) and who has fraudulently concocted the numbers accordingly, digital development analysis is performed in order to unmask such hidden and sophisticated fraud. It is assumed that such a well-educated cheater would nonetheless be unaware of digital development pattern and thus invent numbers with either the steady logarithmic condition throughout the entire data, or with a meaningless zigzag style. The table in Fig. 3.18 depicts digital configuration for each sub-interval between the relevant IPOT points. The correct transition from approximately digital equality on the left to extreme skewness on the right can be clearly confirmed in the table, yet a more rigorous analysis shall be performed examining the formal results of the four developmental tests outlined in the previous chapter.

The ES12 vector of digital development along the entire range of data is:  $\{-30.5\%, -16.2\%, -0.8\%, +3.3\%, +28.4\%, +52.3\%\}$ . The first sub-interval on the extreme left of (0.01, 0.10) contains only 13 points representing a mere 0.09% portion of overall data and thus it is considered an outlier not to be included in all subsequent calculations. Simple Linear Regression is performed for ES12 vector of development — thought of as the dependent variable — which is regressed on  $\{1, 2, 3, 4, 5, 6\}$  serving as the vector of independent variable  $N_i$ .

<b>Left Border Point</b>	<b>0.1</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>
<b>Right Border Point</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>	<b>100,000</b>
	===	===	====	=====	=====	=====
<b>Digit 1</b>	4.4	16.7	29.3	31.6	56.6	69.4
<b>Digit 2</b>	12.8	14.8	17.6	19.5	19.5	30.6
<b>Digit 3</b>	24.9	17.7	12.7	13.8	9.5	0.0
<b>Digit 4</b>	8.6	10.0	9.5	9.4	5.6	0.0
<b>Digit 5</b>	7.9	10.6	8.2	7.6	3.5	0.0
<b>Digit 6</b>	8.2	8.2	8.2	5.9	2.4	0.0
<b>Digit 7</b>	8.6	7.1	5.5	4.5	1.5	0.0
<b>Digit 8</b>	15.8	6.2	4.6	4.6	1.2	0.0
<b>Digit 9</b>	8.9	8.7	4.4	3.1	0.5	0.0
	-----	-----	-----	-----	-----	-----
<b>Data points:</b>	571	2910	5275	5290	1099	36
<b>% Overall Data</b>	3.8%	19.2%	34.7%	34.8%	7.2%	0.2%
<b>ES12</b>	-31%	-16%	-1%	3%	28%	52%

Figure 3.18 Digital Development Pattern — Canford Audio PLC Price List



**Figure 3.19** ES12 Developmental Line — Canford Audio PLC Price List

Performing linear regression, we obtain:  $ES12_i = +15.8 * Ni - 49.1$ , namely  $-49.1\%$  intercept and  $+15.8\%$  slope. Figure 3.19 depicts those six ES12 points in black color, as well as the fitted regression line in gray color. Substituting  $N = 1$  and  $N = 6$  for the two extreme sub-intervals on the leftmost and rightmost sides, we obtain  $(15.8\% * 1 - 49.1\%)$  and  $(15.8\% * 6 - 49.1\%)$  respectively, therefore:

$$ES12(\text{left}) = -33.3\% \quad \text{and} \quad ES12(\text{right}) = +45.7\%$$

Hence Canford Price List easily passes developmental Test I, since

$$ES12(\text{left}) < -25\% \quad \text{and} \quad ES12(\text{right}) > +25\%$$

Calculating sum of absolute values of all six ES12 values, we obtain:

$$|-30.5\%| + |-16.2\%| + |-0.8\%| + |+3.3\%| + |+28.4\%| + |+52.3\%| = 131.5\%, \text{ and thus Canford Price List easily passes developmental Test II:}$$

Sum of absolute values of all ES12  $> 110\%$ .

Examination of digital development along the second-order line also confirms the expected pattern that should be found in all random data types. The vector of ES04 is:

$$ES04 = \{-10.9\%, -1.2\%, -1.2\%, +1.5\%, +8.8\%, +34.2\%\}$$

Performing Simple Linear Regression, we obtain  $ES04_i = +7.4 * Ni - 20.6$ .

Substituting  $N = 1$  and  $N = 6$  for the two extreme sub-intervals on the leftmost and rightmost sides, we obtain  $(7.4\% * 1 - 20.6\%)$  and  $(7.4\% * 6 - 20.6\%)$  respectively, thus:

$$ES04(\text{left}) = -13.2\% \quad \text{and} \quad ES04(\text{right}) = +23.6\%$$

Hence Canford Price List easily passes Test III, since

$$ES04(\text{left}) < -2\% \quad \text{and} \quad ES04(\text{right}) > +6\%$$

Canford Price List also easily passes Test IV, since

$$|-10.9\%| + |-1.2\%| + |-1.2\%| + |+1.5\%| + |+8.8\%| + |+34.2\%| = 57.9\%$$

Sum of absolute values of all ES04 > 15%

With regard to summation test in the context of fraud detection, one would certainly expect to obtain here the usual sixfold approximate advantage for digit 1 as compared with digit 9. This is so not only due to the fact that revenue data as well as price lists are of the random flavour and thus must show the usual disparity among the digits in amount of sum obtained, but also more specifically so here since digital development was explicitly shown to exist for Canford price list data.

**It is digital development that drives the disparity in sums along digital lines!** Figures 3.20 and 3.21 indeed confirm that sum distributions are strongly skewed in favor of low digits and low digit combinations, with approximately six-fold first-digit advantage, and ninefold FTD advantage. The spike at 24 on the FTD summation chart should not raise any suspicion since this is simply due to the fact

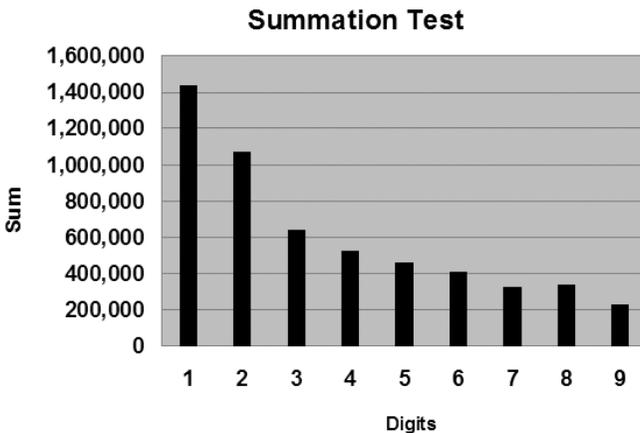


Figure 3.20 Sums Along First Digits — Canford Audio PLC Price List

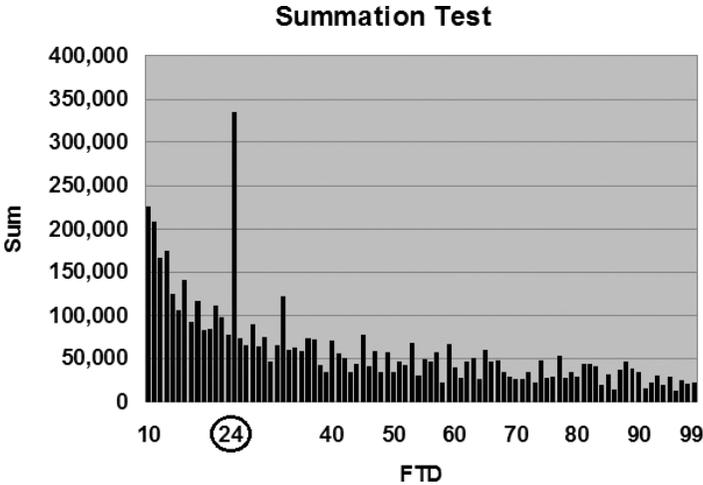


Figure 3.21 Sums Along First-Two Digits — Canford Audio PLC Price List

Value	Frequency
3.40	149
0.36	67
16.67	67
1.65	62
69.30	56
9.11	48
14.22	46
17.87	44
32.92	41
0.83	39

Figure 3.22 Value Repetition Test — Canford PLC

that upper end of data abruptly terminates in values around \$24,198.90 and \$24,896.85 for some of the most expensive items in the price list.

Finally, value repetition test is performed as depicted in Fig. 3.22, showing the ten-most-frequent values in the price list. The value \$3.40 is the most repeated one in the price list, appearing 149 times. This nicely explains why the highest deviation in the first order occurs for digit 3 which obtained 14.2% leadership here instead of the expected 12.5% share allocated by the law. Had the data been truly about revenue amounts then a slight suspicion would have been raised about \$3.40 being perhaps repeatedly invented in the mind of a scheming accountant,

prompting further auditing of all invoices with amount of \$3.40. Since the data represents the price list of the company, no suspicion arises here, as there are simply many items for sale costing that amount.

In conclusion, treating Canford price list as revenue data, nothing improper is found here contrary to BL to cause any suspicion whatsoever, except the six spikes on the LTD chart. When the data is acknowledged as price list and the reasons for those LTD spikes are then explained, its agreement with the logarithmic is deemed to be highly satisfactory.

What should the data analyst expect for actual Canford revenue data in a digital sense? Given the enormous size of the price list, and that it is almost perfectly logarithmic in and of itself, any Random Linear Combination from it as in any typical invoice should show even more conformity to Benford's Law than the price list itself. RLC simulation scheme with 8000 runs, assuming customers purchase exactly one item, with quantities chosen randomly as in the discrete Uniform {1, 2, 3}, yields the following results:

Random Linear Combination = [Canford Price List]\*Uniform {1, 2, 3}

1st digits — {30.8, 16.9, 12.4, 8.9, 7.4, 7.3, 6.3, 5.1, 4.9}, SSD = 2.6

2nd digits — {12.5, 10.8, 11.9, 10.8, 9.6, 9.5, 9.8, 8.6, 8.1, 8.6}, SSD = 2.9

Such low SSD values confirm the above expectation that revenue data should come close to perfection in logarithmic behavior. Therefore, any future Canford revenue data deviating even mildly from BL should raise suspicion and prompt comprehensive audit.

## CASE STUDY IX: USA COUNTY AREA DATA

---



---

Forensic digital analysis shall be performed on U.S. County Area data. The data pertains to areas of all 3,143 counties in the USA. Data can be downloaded from the U.S. Census website <http://www.census.gov/support/USACdataDownloads.html#LND> where “Land Area” is selected with the choice of LND01.xls for data downloads. The data on column X called LND110210D is selected for analysis. It is necessary to eliminate from the data the following items: (a) 50 aggregated values of 50 states, (b) area for District of Columbia, (c) total area of USA, (d) three entries with zero as area; resulting in areas for 3,143 proper counties. Data is reported in units of **Square Miles**. Some examples are shown below:

Yukon-Koyukuk County in Alaska has the largest area — 145,504.79 sqr miles  
 North Slope County in Alaska has the second largest area — 88,695.41 sqr miles  
 Otoe County in Nebraska has the median area size — 615.63 sqr miles  
 Falls Church County in Virginia has the second smallest area — 2.50 sqr miles  
 Lexington County in Virginia has the smallest area — 2.00 sqr miles

Certainly in this case, fraud detection is not an issue. The U.S. Census Office is trusted in providing honest data, yet this case is quite instructive and useful in demonstrating many relevant procedures and ideas in forensic digital analysis. Analyses of the following five basic digital distributions shall be performed:

The 1st order digit distribution is:

U.S. County — {16.2, 10.0, 10.7, 15.8, 15.2, 10.4, 8.6, 7.1, 5.9}  
 BL 1st Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

The 2nd order digit distribution is:

U.S. County — {13.5, 11.1, 10.2, 10.0, 8.5, 9.3, 9.4, 9.8, 8.6, 9.7}  
 BL 2nd Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

The 3rd order digit distribution is:

U.S. County — {10.40, 10.44, 10.15, 9.74, 9.86, 9.55, 9.74, 10.24, 9.86, 10.02}

BL 3rd Digits — {10.18, 10.14, 10.10, 10.06, 10.02, 9.98, 9.94, 9.90, 9.86, 9.83}

FTD chart depicted in Fig. 3.23 is clearly non-logarithmic. LTD chart depicted in Fig. 3.24, on the other hand, appears nicely logarithmic; it is just as compatible with the law as are most other LTD charts of logarithmic data sets. This fact is also confirmed by its 3.8 SSD value, which is below the arbitrary cutoff value of 4 in the table of Fig. 3.3, classifying it as ‘perfectly Benford’. [Note: 16 values between 2.00 and 9.88 were omitted for the LTD chart since these numbers are too short digit-wise and squarely belong to the first, second, and third orders].

SSD of 1st digits yields the very high value of 371.8

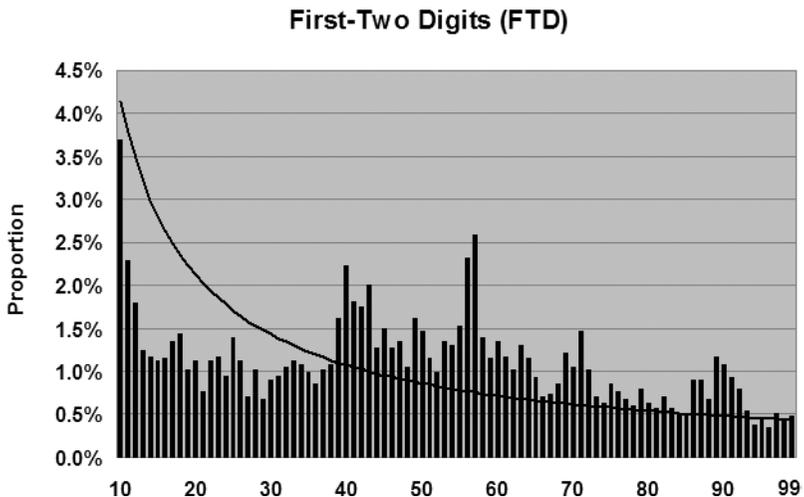
SSD of 2nd digits yields the low value of 7.7

SSD of 3rd digits yields the extremely low value of 0.66

SSD of FTD yields the moderately high value of 46.0

SSD of LTD yields the very low value of 3.8

All in all, U.S. County Area data set certainly failed the logarithmic test. **The fact that higher-orders digit distributions do indeed behave highly logarithmically does not excuse the data set in its entirety, since such**



**Figure 3.23** First-Two Digits for U.S. County Area Data set

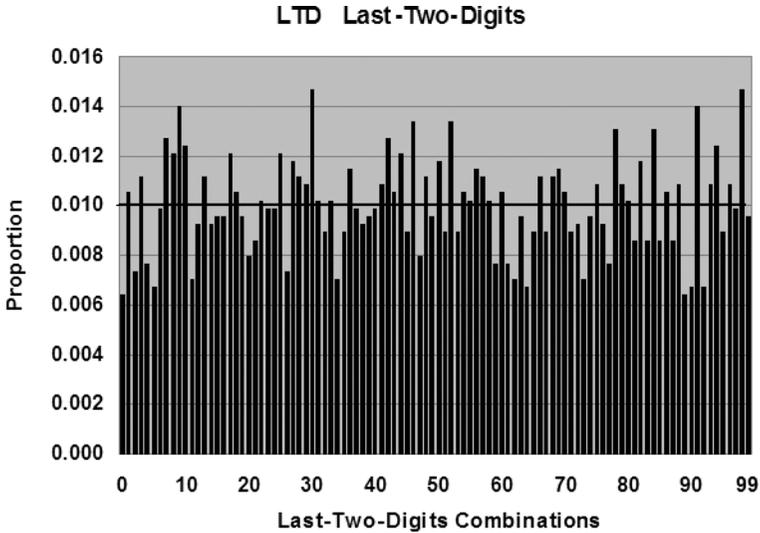


Figure 3.24 Last-Two Digits for U.S. County Area Data set

**selective compliance with the law is extremely typical in most non-logarithmic data types.** Chapter 71 titled ‘The Near Indestructibility of Higher Order Distributions’ shall examine the theoretical basis for such odd compliance with the law for most data sets, logarithmic or non-logarithmic. The conceptual understanding for the digital behavior of U.S. County Area data is that this data set is too narrowly focused on a limited range of values without sufficient order of magnitude. County boundaries within each state are man-made and artificial; legal declarations done in a way so as to give approximately similar geographic areas to each county. State officials would rarely assign counties huge land areas (at the expense of others), nor would they usually assign any given county tiny area. Therefore the spread of the data is too small, focused narrowly on certain magnitudes. The extreme cases shown above [Alaska and Virginia] with deceptive high order of magnitude (of supposedly about 4.9) are merely four examples at the extreme ends of the data set, truly constituting outliers on the margin of data, while the bulk of the data is spread over a much narrower range. Applying the general logarithmic rule of Order of Magnitude of Variability discussed in Chapter 10, namely:

OMV = LOG(90th percentile) – LOG(10th percentile) > 3, we obtain:

90th percentile = 1843.1 → LOG(1843.1) = 3.27

10th percentile = 286.0 → LOG(286.0) = 2.46

OMV = LOG(90th percentile) – LOG(10th percentile) = 3.27 – 2.46 = 0.81 << 3

This clearly explains why U.S. County Area data set is not logarithmic at all.

Converting the data set into a scale of **Square Kilometers**, using the conversion formula *1.00 Square Mile = 2.5899 Square Kilometer*, the five examples above become:

Yukon-Koyukuk County in Alaska has the largest area — 376,856.0 sqr km

North Slope County in Alaska has the second-largest area — 229,720.2 sqr km

Otoe County in Nebraska has the median area size — 1594.5 sqr km

Falls Church County in Virginia has the second-smallest area — 6.5 sqr km

Lexington County in Virginia has the smallest area — 5.2 sqr km

For U.S. County Area data quoted in square kilometers, we obtain the following results for the first-, second- and third-orders digits:

1st Digits: {50.0, 20.3, 5.6, 4.8, 3.8, 4.5, 3.1, 4.2, 3.8}, SSD = 503.4

2nd Digits: {12.3, 11.2, 11.5, 11.5, 12.0, 9.3, 8.8, 8.0, 8.6, 6.9}, SSD = 9.6

3rd Digits: {9.90, 9.80, 9.55, 10.47, 9.96, 9.45, 10.63, 10.85, 9.86, 9.55}, SSD = 2.40

Clearly, drastic changes for the first-digit configuration take place here due to the change in scale. It should be quite instructive to see how exactly this change is occurring, and this is nicely accomplished via the examination of the median.

Otoe County in Nebraska has the median area size — 615.63 square miles

Otoe County in Nebraska has the median area size — 1594.50 square kilometers

Since the first digit of the median value and all the numerous values around it change from 6 to 1 under the mile-to-kilometer scale transformation, the effect on first order is to strongly increase digit 1 leadership, and this is why the new proportion for digit 1 is 50.0% as compared with the original 16.2% proportion. The scale invariance principle cannot be applied here since the data set is not logarithmic to begin with, and this is why first digits are dramatically affected by scale changes. Yet again, higher orders show their resilience and tenacity, as can be seen here indeed, where second- and third-digit orders are not being adversely affected by the scale change for the most part. It should also be noted that the scale change did not alter OMV.

$$90\text{th percentile} = 4773.5 \rightarrow \text{LOG}(4773.5) = 3.68$$

$$10\text{th percentile} = 740.7 \rightarrow \text{LOG}(740.7) = 2.87$$

$$\text{OMV} = \text{LOG}(90\text{th percentile}) - \text{LOG}(10\text{th percentile}) = 3.68 - 2.87 = 0.81$$

Interestingly, even though U.S. County Area data is decisively non-logarithmic, nonetheless it shows a clear digital development pattern, just as is found in all random logarithmic data sets! Figure 4.76 of Chapter 84 in the fourth section depicts this development pattern in more detail. Calculations show that ES12 sequence along the partition  $\{1, 10, 100, 1000, 10000, 100000\}$  yields the vector of development  $\{-29.0\%, -8.7\%, -36.1\%, +29.7\%, +34.1\%\}$ . The last sub-interval on the extreme right of (100000, 1000000) contains very little data [only 0.03% portion of overall data, having only one case] and thus it is considered an outlier not to be included in all subsequent calculations. Simple Linear Regression is performed for ES12 vector of development — thought of as the dependent variable — which is regressed on  $\{1, 2, 3, 4, 5\}$  serving as the vector of independent variable  $N_i$ . Performing linear regression, we obtain:  $\text{ES12}_i = +16.5 * N_i + -51.3$ , namely  $-51.3\%$  intercept and  $+16.5\%$  slope. Substituting  $N = 1$  and  $N = 5$  for the two extreme sub-intervals on the leftmost and rightmost sides, we obtain:

$$\text{ES12}(\text{left}) = -34.9\% \quad \text{and} \quad \text{ES12}(\text{right}) = +30.9\%$$

Hence US County Area data set easily passes developmental Test I since

$$\text{ES12}(\text{left}) < -25\% \quad \text{and} \quad \text{ES12}(\text{right}) > +25\%$$

Calculating sum of absolute values of all five ES12 values, we obtain:  $|-29.0\%| + |-8.7\%| + |-36.1\%| + |+29.7\%| + |+34.1\%| = 137.6\%$ , and thus U.S. County Area data set easily passes developmental Test II: Sum of absolute values of all ES12  $> 110\%$ .

The forensic digital analyst should categorically refuse to even consider an attempt at calculating the value of any supposed ‘chi-square statistic’ for this data set! Unless he or she can be miraculously convinced that this data set on areas of all the counties in the USA is a truly ‘random sample’ taken from a much larger population of parental (and logarithmic!) data on either global or galactic-wide records on area of counties, cantons, prefectures, regions, districts, states, planets, and star systems [which just by pure chance happened to incorporate values only from USA of Earth and somehow also happened to cover all the counties within the USA; all the while not even a single value from any other country or planet fell into the collection of number in this ‘truly random sampling project’ on areas].

## RANDOM LINEAR COMBINATIONS AND REVENUE DATA REVISITED

---

Chapter 16 on Random Linear Combinations analyzed potential revenue data of one particular small shop with exactly nine items on sale. In this chapter many other cases and shops shall be analyzed and a general rule about logarithmic behavior shall be given.

The following analysis is an example of a particular computer simulation result relating to random linear combinations of one hypothetical very small shop that has only six items for sale. All six items are assumed here equally likely to be sold (equal popularity). The shopper is assumed to purchase exactly two (distinct or identical) items. The shopper is then to roll two dice (one for each item) as a way to decide on the number of units (quantities) he or she may wish to buy — out of {1, 2, 3, 4, 5, 6} possibilities. At most, when both dice show side 6, the shopper will buy 12 units in total. At a minimum, when both dice show side 1, only two units would be purchased (one unit per item). This particular list shall be called ‘original’. The expression of the schematic arrangement of these simulations, the list of prices for the six items for sale, the resultant first leading digits distribution, and its SSD measure, are as follows:

Random Linear Combination = List\*dice1 + List\*dice2

List (**original**) = {\$2.25, \$4.75, \$7.75, \$9.50, \$10.25, \$35.00}

LD = {23.1, 17.4, 11.6, 11.3, 10.5, 9.7, 8.0, 4.6, 3.7} **SSD = 74.0**

The table in Fig. 3.25 depicts a small section from the random output of these computer simulations. [Note: for this simulation, and in all subsequent simulations in this chapter, 20,000 simulated runs will be performed for each scenario/model].

While digits distribution is not quite logarithmic here, it resembles the logarithmic a great deal, and it is monotonically decreasing. This is so in spite of the very low number of products on sale in this very small shop. If we now increase the number of products in the shop to nine by adding three new products valued at \$3.25, \$25.00, and \$37.00, then resultant leading digits distribution gets much

Price for item	Quantity	Sub-total	Price for item	Quantity	Sub-total	Final Bill
\$ 9.50	4	\$ 38.00	\$ 2.25	2	\$ 4.50	\$ 42.50
\$ 7.75	3	\$ 23.25	\$ 35.00	4	\$ 140.00	\$ 163.25
\$ 7.75	3	\$ 23.25	\$ 10.25	1	\$ 10.25	\$ 33.50
\$ 35.00	6	\$ 210.00	\$ 7.75	6	\$ 46.50	\$ 256.50
\$ 35.00	1	\$ 35.00	\$ 9.50	5	\$ 47.50	\$ 82.50
\$ 4.75	5	\$ 23.75	\$ 7.75	2	\$ 15.50	\$ 39.25
\$ 35.00	4	\$ 140.00	\$ 2.25	2	\$ 4.50	\$ 144.50
\$ 9.50	5	\$ 47.50	\$ 4.75	4	\$ 19.00	\$ 66.50
\$ 35.00	1	\$ 35.00	\$ 9.50	5	\$ 47.50	\$ 82.50
\$ 10.25	1	\$ 10.25	\$ 9.50	1	\$ 9.50	\$ 19.75
\$ 7.75	3	\$ 23.25	\$ 10.25	5	\$ 51.25	\$ 74.50
\$ 9.50	2	\$ 19.00	\$ 2.25	2	\$ 4.50	\$ 23.50
\$ 2.25	2	\$ 4.50	\$ 9.50	1	\$ 9.50	\$ 14.00
\$ 10.25	4	\$ 41.00	\$ 4.75	3	\$ 14.25	\$ 55.25
\$ 4.75	3	\$ 14.25	\$ 2.25	6	\$ 13.50	\$ 27.75
\$ 9.50	1	\$ 9.50	\$ 10.25	4	\$ 41.00	\$ 50.50
\$ 10.25	1	\$ 10.25	\$ 9.50	1	\$ 9.50	\$ 19.75

Figure 3.25 Revenue Data As an Example of Random Linear Combinations (original)

closer to the logarithmic, and SSD falls significantly from 74.0 to 16.6. It is noted that the addition of these three new items leaves overall range almost unaltered, from around \$2 to around \$35 or \$37. In other words, by adding three items for sale we have not changed overall price range, only the number of prices. This price list shall be called ‘enlarged’, and it is identical to the example presented in the earlier chapter (16) on Random Linear Combinations. The scheme, the new enlarged list, its resultant leading digits distribution of simulations, and SSD, are as follows:

$$\text{Random Linear Combination} = \text{List} \cdot \text{dice1} + \text{List} \cdot \text{dice2}$$

$$\text{List (enlarged)} = \{ \$2.25, \$3.25, \$4.75, \$7.75, \$9.50, \$10.25, \$25.00, \$35.00, \$37.00 \}$$

$$\text{LD} = \{ 31.5, 20.3, 10.5, 8.6, 8.2, 6.7, 5.9, 5.2, 3.1 \} \quad \text{SSD} = 16.6$$

This demonstrates the importance of having a large number of values within the list of prices in order to obtain better logarithmic behavior.

On the other hand, shrinking our original list to just three items would considerably reduce its variability, and convergence to the logarithmic would become almost impossible. Simulation using one such shorter list of three products clearly shows a considerable deterioration in leading-digit distribution, even though overall range has not changed at all. The scheme, the list, the resultant digital distribution, and SSD are as follows:

Random Linear Combination = List\*dice1 + List\*dice2

List (**reduced**) = {\$2.25, \$4.75, \$35.00}

LD = {38.2, 26.5, 14.6, 7.2, 2.1, 1.3, 4.8, 2.5, 2.8} SSD = 229.2

Another consideration is the overall variability of the values being combined — in terms of width of range — measured as the money difference between the most expensive item and the cheapest item. It is necessary to have sufficiently large overall range of values in the list of prices in order to obtain good logarithmic behavior. The result of another simulation is shown pertaining to a scheme similar to the original one given above, but with a list of six prices that are too narrowly focused around \$20 without much spread. Such a scheme, the list, its resultant leading digits distribution, and SSD, are as follows:

Random Linear Combination = List\*dice1 + List\*dice2

List (**narrowly focused**) = {\$17.75, \$18.00, \$20.50, \$21.25, \$25.00, \$26.50}

LD = {61.0, 19.1, 1.2, 1.7, 1.9, 2.5, 4.6, 3.6, 4.5}

SSD = 1205.8

This result demonstrates that without sufficient spread in the set of prices (i.e. sufficient order of magnitude), random linear combination would not lead to any logarithmic convergence whatsoever. Moreover, having even thousands or millions of items on sale, and not merely six, all falling on a very narrow price range, would result in similar lack of logarithmic convergence.

In general, it is necessary to keep all aspects as random as possible in order to get enough spread and variability in the resultant linear combinations, so that logarithmic convergence is obtained. For example, let us re-consider the original case of a small shop with six items on its list, and assume one peculiar class of shoppers that for some reason always decide to buy all six available types of products, while rolling six dice to decide upon the number of units each. Here, we have totally removed the randomness and uncertainty regarding the type of items to be bought, since all items are certain to be bought in such a scheme, with the result

that leading digits are not even remotely logarithmic. The scheme, the list, the resultant digital distribution, and SSD, are as follows:

$$\begin{aligned} \text{Random L.C.} &= 2.25*\text{dice1} + 4.75*\text{dice2} + 7.75*\text{dice3} + 9.5*\text{dice4} \\ &\quad + 10.25*\text{dice5} + 35.00*\text{dice6} \\ \text{List (original)} &= \{\$2.25, \$4.75, \$7.75, \$9.50, \$10.25, \$35.00\} \\ \text{LD} &= \{29.0, 47.5, 23.1, 0.1, 0.0, 0.0, 0.0, 0.1, 0.3\} \quad \mathbf{SSD = 1283.7} \end{aligned}$$

To demonstrate the power of random linear combination in general, another particular result is shown here where the logarithmic is nearly achieved with a mere six-product list:

$$\begin{aligned} \text{Random Linear Combination} &= \text{List}*\text{dice1} + \text{List}*\text{dice2} \\ \text{List (quick convergence)} &= \{\$1.25, \$1.75, \$6.75, \$12.50, \$35, \$58\} \\ \text{LD} &= \{30.8, 19.6, 14.0, 9.7, 5.5, 6.6, 6.2, 5.1, 2.6\} \\ &\quad \mathbf{SSD = 16.6} \end{aligned}$$

It is quite remarkable that random combinations from a list having just six items come out so close to the logarithmic! On the other hand, for the above list of 'quick convergence', if one assumes that the shopper always buys just one item (reducing randomness), and rolling a dice to decide on the number of units, then resultant digits distribution is not as logarithmic; rather, it's a bit off. The model, the list, the resultant leading digits distribution, and SSD, are as follows:

$$\begin{aligned} \text{Random Linear Combination} &= \text{List}*\text{dice} \\ \text{List (quick convergence)} &= \{\$1.25, \$1.75, \$6.75, \$12.50, \$35, \$58\} \\ \text{LD} &= \{28.3, 19.3, 16.5, 2.8, 11.1, 8.6, 10.9, 2.6, 0.0\} \\ &\quad \mathbf{SSD = 136.7} \end{aligned}$$

In another simulation, the number of items itself was kept random, assigning 50% probability that the shopper would purchase one item, 30% that two items are purchased, and 20% that three items are purchased. Such an arrangement biases in general the total bill towards low values which abound, while large values occur more rarely. Such simulation arrangement often serves as a more realistic model, more appropriate and fitting for the real-life case of the shopper who tends to buy fewer items more often than numerous ones. The model, the list, the resultant leading digits distribution, and SSD are as follows:

$$\begin{aligned} \text{Random Linear Combination} &= \\ 50\% &: [\text{List}*\text{dice}] \\ 30\% &: [\text{List}*\text{dice1} + \text{List}*\text{dice2}] \\ 20\% &: [\text{List}*\text{dice1} + \text{List}*\text{dice2} + \text{List}*\text{dice3}] \end{aligned}$$

List (**original**) = {\$2.25, \$4.75, \$7.75, \$9.50, \$10.25, \$35.00}

LD = {26.4, 18.5, 12.0, 11.7, 7.4, 7.3, 6.8, 3.0, 6.8} **SSD = 29.7**

Indeed, digital results here for the original price list are much closer to the logarithmic in this scenario than under the assumption that exactly two items are being purchased seen at the beginning of this chapter with its higher SSD value of 74.0.

Is it possible to make some generalizations and predictions regarding resultant leading digits distributions of models of random linear combinations from the setup of the process itself? Are there any hard and fast rules? A lot seems to depend on the particular values within the price list, the spread of its range, the leading digits distribution of the price list itself, as well as its numerical size, namely the number of values within the list. Needless to say, the specification within the model regarding how many items are being picked is also an important factor, as well as the manner or scheme deciding on the quantity chosen from each item, which may be done via a fair dice of six sides, or in any other conceivable way.

In another example, we allow a continuous range to serve as the basis (approximation) for the discrete price list. If the histogram of a very long price list of a very large retail corporation such as, say, Walmart with numerous items for sale, closely resembles the Uniform distribution, then simulated values from the Uniform chosen to represent the price, multiplied by, say, the value of a simulated dice {1, 2, 3, 4, 5, 6} representing quantity bought, could serve as the model for the shopper who buys only one type of product at Walmart. In reality, typical price lists of large retail stores are never as in the Uniform distribution, nor as in the Normal distribution. Rather, they are more like, say, the exponential or Lognormal distributions, which are heavily skewed in favor of low quantities (having numerous cheap products, some reasonably priced items in the middle, and very few truly expensive items). In any case, results from computer simulations of the model **Uniform(0, Upper Bound)\*dice** depend on the value assigned to upper bound, but are in general quite close to the logarithmic. A summary of output from such 20,000 computer simulations each is shown in Fig. 3.26.

Such results are quite remarkable, given that the Uniform itself is not logarithmic at all. In fact, it may be considered in a sense ‘un-logarithmic’ or ‘anti-logarithmic’, and especially so when defined between two adjacent IPOT points such as (10, 100) for example where all digits are equally likely to occur. Driving the

Upper Bound:	100	200	300	400	500	600	700	800	900
Digit 1	29.7	25.9	30.5	34.1	36.7	36.1	33.9	32.3	30.6
Digit 2	19.5	14.0	16.1	13.8	17.2	19.8	21.4	21.2	21.0
Digit 3	14.3	14.4	9.0	11.9	9.0	10.0	12.7	14.6	14.4
Digit 4	10.7	10.4	10.1	6.9	9.1	7.0	7.5	8.6	9.9
Digit 5	7.2	10.1	8.9	7.5	6.1	8.3	5.9	4.8	6.4
Digit 6	4.8	7.3	6.7	7.5	5.7	3.9	6.4	5.6	4.9
Digit 7	4.9	7.8	7.4	7.8	5.6	5.0	4.0	5.5	5.1
Digit 8	3.8	6.2	6.7	5.2	4.9	5.2	3.8	3.6	4.8
Digit 9	5.2	3.9	4.8	5.3	5.6	4.8	4.5	3.8	2.9

Figure 3.26 Digital Results of RLC Model Uniform(0, Upper Bound)\*dice

Upper Bound:	100	200	300	400	500	600	700	800	900
Digit 1	18.5	29.2	43.3	42.8	37.0	32.3	28.1	24.1	21.3
Digit 2	19.6	8.8	11.7	21.8	26.5	26.7	24.9	23.4	22.1
Digit 3	17.9	9.8	6.1	7.3	11.9	16.6	19.4	19.6	18.8
Digit 4	15.5	9.9	6.3	4.7	5.6	8.1	11.0	13.6	14.1
Digit 5	11.2	9.4	6.6	4.4	3.7	4.0	5.6	8.0	10.1
Digit 6	7.6	9.6	6.7	4.6	3.7	3.2	3.4	4.3	5.8
Digit 7	4.6	8.7	6.4	4.8	3.4	3.1	2.6	2.8	3.4
Digit 8	3.0	7.7	6.6	4.9	4.0	3.0	2.5	2.2	2.3
Digit 9	2.3	6.7	6.2	4.8	4.2	3.1	2.5	2.1	2.1

Figure 3.27 Digital Results of RLC Model U(0, UB)\*dice1 + U(0, UB)\*dice2

above result though are two levels of randomness, the first being the particular value chosen randomly from the Uniform (the product chosen), and the second being the dice thrown randomly choosing a multiplicative factor (the quantity bought). Interestingly, a linear combination of two Uniform values, representing the type of shoppers who are determined to buy exactly two products, and where the quantity of each product is determined by a roll of a dice, namely the model **Uniform(0, Upper Bound)\*dice1 + Uniform(0, Upper Bound)\*dice2**, yields worsening digital results as shown in Fig. 3.27.

And even worse results are gotten for the triple random linear combination of Uniforms representing the type of shoppers buying exactly 3 items, namely the model **Uniform(0, UB)\*dice1 + Uniform(0, UB)\*dice2 + Uniform(0, UB)\*dice3**.

In one simulation run, with  $UB = 100$ , first digits distribution is  $\{10.5, 11.8, 14.1, 15.7, 14.8, 12.9, 9.5, 6.8, 3.9\}$ .

Even if one assumes that the typical shopper has  $1/3$  chance of buying one item [from the uniformly distributed price list],  $1/3$  chance of buying two items, and  $1/3$  chance of buying three items [in other words, when the above three RLC schemes are mixed in equal proportions], overall leading digits are still not logarithmic. In one such scenario, and where  $UB$  is 100, this more complex scheme yields  $\{19.2, 17.3, 15.9, 13.6, 11.5, 8.1, 6.1, 4.7, 3.7\}$ ,  $SSD = 161.7$ .

The power of random linear combinations is elegantly illustrated though in the case of the exponential distribution. The exponential distribution is close to being logarithmic in its own right, but it is never close enough, and this is so regardless of parametrical value. Yet these mild deviations from the logarithmic can be easily corrected and repaired by way of random linear combinations. No matter what parameter is chosen, the model **exponential(parameter)\*dice** is extremely close to the logarithmic! Such decisive improvement in digital behavior is achieved simply by the random multiplicative factor of the dice!

It is noted that a definite deterioration sets in for the random linear combinations **exponential(parameter)\*dice1 + exponential(parameter)\*dice2**, and even worse results are gotten for the triple combination, and so forth, just as was seen in the case of the Uniform distribution above.

**A priori** there are no hard and fast rules as to logarithmic success. Yet in retrospect, there exists a general criteria that might take most if not all the mystery out of the digital phenomenon of Random Linear Combinations. The **posteriori** criteria that could serve almost as a general rule, indicating whether a given RLC process yields or does not yield the logarithmic, is the consideration of the resultant **Order Of Magnitude (OOM)**. We have defined order of magnitude as **LOG[(Max/Min)] = LOG[Max] - LOG[Min]**, where Min stands for the lowest value, and Max for the highest value, of resultant or generated data (not of the price list itself). For example, resultant data that is spread within the range of  $[10, 10000]$  has order of magnitude  $LOG[10000/10] = LOG[1000] = 3$ , and roughly speaking, it spans three regions standing in between adjacent IPOT points, namely the regions  $(10, 100)$ ,  $(100, 1000)$ , and  $(1000, 10000)$ . Surely, each random linear combination process points to a unique order of magnitude of resultant data. It would do no good of course in the definitions of Min and Max to consider outliers within the data, those extreme points very far away from the main body of values where almost all the data lies. Instead, the approximate edges of the main

body of data should serve as Min and Max values. It must be carefully noted and emphasized here that we are not interested **directly** in the OOM of the price list, but rather in the OOM of the resultant/generated data. Yet, a large value of OOM for the price list often implies also a large value of OOM in the resultant data for a given structure of the model; hence **indirectly** we are indeed interested in the OOM value of the price list, as it generally provides a good indication of logarithmic convergence. For example, OOM of the ‘original’ price list {\$2.25, \$4.75, \$7.75, \$9.50, \$10.25, \$35.00} is  $\text{LOG}(35.00/2.25)$  or 1.2, while OOM of output data for the Random Linear Combination model **‘original’\*dice1 + ‘original’\*dice2** has a bit larger value and it is calculated as  $\text{LOG}(2*6*35.00/2*1*2.25)$  or **1.97**, namely the log distance between two large and excessive purchases of six quantities of the most expensive item, and two frugal purchases of one quantity of the cheapest item on the list. For the ‘narrowly focused’ price list of {\$17.75, \$18.00, \$20.50, \$21.25, \$25.00, \$26.50}, OOM value of the price list itself is  $\text{LOG}(26.50/17.75)$  or 0.2, while OOM value of output data relating to the RLC model **‘narrowly focused’\*dice1 + ‘narrowly focused’\*dice2** is  $\text{LOG}(2*6*26.50/2*1*17.75)$  or **0.95**. Hence the difference in logarithmic behavior between ‘original’ and ‘narrowly focused’ is easily explained by way of their different OOM values! RLC output of ‘narrowly focused’ is not logarithmic at all because its OOM value is 0.95 which is too narrow, while RLC output of ‘original’ is much closer to the logarithmic because its OOM value of 1.97 is much larger.

As shall be seen in later chapters in the fourth section, by and large an order of magnitude larger than 2.5 or 3 is normally sufficient for any quantitatively-skewed data set to behave approximately logarithmically, unless the data is too discrete with frequent and huge gaps between values, not resembling anything continuous, or unless the data artificially starts/stops at a fixed value abruptly without any graduation. Data sets restricted to orders of magnitude of less than two are rarely logarithmic. For example, adult human weight is roughly restricted to [30, 250] kilograms, having an order of magnitude  $\text{LOG}[250/30] = \text{LOG}[8.33] = 0.92$ , and such low OOM value is insufficient for logarithmic behavior.

Furthermore, **order of magnitude is scale invariant!** If data is spread within the range of  $(10^R, 10^Q)$ , then it has  $\text{LOG}(10^Q/10^R) = \text{LOG}(10^Q) - \text{LOG}(10^R) = Q*\text{LOG}(10) - R*\text{LOG}(10) = (Q - R)$  order of magnitude. Any re-scaling of the data by the same factor F applied to all data points implies that the transformed data is now being spread within the new range of  $(F*10^R, F*10^Q)$ , or  $(10^{\text{LOG}(F)} * 10^R, 10^{\text{LOG}(F)} * 10^Q)$ , namely over the range of  $(10^{R+\text{LOG}(F)}, 10^{Q+\text{LOG}(F)})$ ,

and order of magnitude of re-scaled data is completely unchanged at  $[(Q + \text{LOG}(F)) - (R + \text{LOG}(F))] = (Q - R)$ .

For example, adult human weight in British Pounds (1 Pound equals 0.4536 Kilograms), varies approximately on the range of [66, 551] and its order of magnitude is  $\text{LOG}[551/66] = \text{LOG}[8.33] = 0.92$ , which is unchanged from our earlier calculation when weights were considered in kilograms. The fact that order of magnitude is scale-invariant gives the measure universality, rendering it a reliable and steady criterion of logarithmic behavior. Yet, it is important to keep in mind that having a large order of magnitude does not guarantee that data is logarithmic, as it is not a totally sufficient condition, even though it plays a very big role in influencing logarithmic behavior, as shall be seen in later chapters. As an important counter example, we examine RLC of a price list having a very high OOM value, and where digits distribution is not logarithmic at all due to congregations of values at the left and at the right poles, leaving the middle part totally empty.

Random Linear Combination = List\*dice1 + List\*dice2

List (**variable but poled**) = {\$2.54, \$3.72, \$4.70, \$8766, \$9451, \$10400}

LD = {17.8, 18.7, 16.2, 12.5, 12.1, 7.1, 4.1, 5.8, 5.7}

**SSD = 196.5**

Calculating OOM for the price list yields  $\text{LOG}(10400/2.54)$  or 3.6, while OOM of output data for the RLC model is  $\text{LOG}(2*6*10400/2*1*2.54)$  or 4.4, since the highest possible linear combination would be when 6 is shown on the two dice and the most expensive item (\$10400) is bought twice, while the lowest possible linear combination would be when 1 is shown on the two dice and the cheapest item (\$2.54) is bought twice.

In another counter example, showing that OOM is not the exclusive factor in logarithmic convergence and does not constitute the final word in Random Linear Combinations, a comparison and re-examination of two price lists and their associated RLC schemes seen earlier illustrate this point clearly:

Random Linear Combination = List\*dice1 + List\*dice2

List (**original**) = {\$2.25, \$4.75, \$7.75, \$9.50, \$10.25, \$35.00}

LD = {23.1, 17.4, 11.6, 11.3, 10.5, 9.7, 8.0, 4.6, 3.7} **SSD = 74.0**

List (**enlarged**) = {\$2.25, \$3.25, \$4.75, \$7.75, \$9.50, \$10.25, \$25.00, \$35.00, \$37.00}

LD = {31.5, 20.3, 10.5, 8.6, 8.2, 6.7, 5.9, 5.2, 3.1} **SSD = 16.6**

Obviously RLC of ‘enlarged’ is much closer to the logarithmic than RLC of ‘original’ is, even though OOMs for both price lists are just about equal here. A better explanation for this divergence in results is that the list ‘enlarged’ is more continuous-like, while the list ‘original’ is more discrete-like.

Having gained sufficient understanding of the main factors leading to logarithmic convergence in RLC processes, let us attempt to apply it to price lists where values are roughly (discretely) uniform [namely, evenly spread within its entire range, where concentrations are equal on the left, central, and right regions]. Two examples, employing the natural numbers and the odd numbers within price lists are shown:

Random Linear Combination = List\*dice1 + List\*dice2

List (**natural and uniform**) = {1, 2, 3, 4, 5, 6, 7, 8, 9}

LD = {17.3, 20.8, 20.4, 14.6, 10.6, 7.9, 3.8, 2.1, 2.5} **SSD = 286.7**

Random Linear Combination = List\*dice1 + List\*dice2

List (**odd and uniform**) = {1, 3, 5, 7, 9, 11, 13, 15, 17}

LD = {23.9, 9.3, 9.1, 13.3, 11.4, 9.5, 8.2, 7.9, 7.4} **SSD = 173.6**

In the case of ‘natural and uniform’, OOM of the generating price list is  $\text{LOG}(9/1)$ , or 0.95, and OOM of its resultant data stands at  $\text{LOG}(2*6*9/2*1*1)$  or 1.73. In the case of ‘odd and uniform’ price list, OOM of the generating price list itself is  $\text{LOG}(17/1)$ , or 1.2, while OOM of its resultant data stands at  $\text{LOG}(2*6*17/2*1*1)$  or 2.0. Hence OOM for either process is not large enough to lead to logarithmic convergence, as seen in the results above. In contrast, in the case of the Uniform distribution, OOM of resultant data is larger. For example, for the RLC model **Uniform(1, 100)\*dice** (where the troublesome number 0 is being substituted by the relatively safe choice of 1), OOM for the RLC model is  $\text{LOG}(6*100/1*1) = \text{LOG}(600) = 2.8$ , which is (almost) sufficiently large.

**A crucial qualification must be made to the Order of Magnitude principle.** Imagine the following digital thought experiment: 99,000 values are chosen from a decisively non-logarithmic data set having the small OOM value of 2, all to be mixed with 1000 values from a nicely logarithmic data set with sufficiently large OOM. Surely the newly created data set of the aggregated 100,000 values is not logarithmic, since Benford’s Law applies to the totality of the data set in question. Nonetheless, OOM of the newly created data is sufficiently large since it contains within it 1% such variable data type.

We therefore arrive at the following conclusion:

**Order of Magnitude should not be applied to the data in its entirety. Rather, OOM must be measured against the bulk of the data, eliminating not only outliers prior to OOM calculations, but also eliminating roughly, say, the lowest 10% and highest 10% of values from the sorted data.** Hence a more refined and effective measure determining logarithmic behavior of data sets in general [not only applicable to RLC models, but also in extreme generality within the whole discipline of Benford' Law] is given by what is called **Order of Magnitude of Variability (OMV)** defined via the general expression:

$$\text{OMV} = \text{LOG}(90\text{th percentile}) - \text{LOG}(10\text{th percentile})$$

$$\text{OMV} = \text{LOG}(X_{90\%}) - \text{LOG}(X_{10\%}).$$

Applying this essential qualification to the OOM principle in the case of RLC model Uniform(0, 100)\*dice, demonstrates clearly the many pitfalls awaiting the novice digital analyst. A direct attempt at OOM calculation yields  $\text{LOG}(100/0)$ , or  $\text{LOG}(\infty)$ , which is  $\infty$ . In order to understand this supposedly peculiar result, one must realize that between 0 and 1 there are infinitely many intervals between adjacent IPOT points, such as (0.1,1), (0.01,0.1), (0.001,0.01), and so forth. Should such 'large' OOM value then automatically guarantee convergence for all RLC models based on the Uniform(0,100)? The answer is decisively in the negative, for the simple reason that most of the data of Uniform(0,100) falls on (1,100), 99% of it to be exact, and having the low OOM value of  $\text{LOG}(100/1)$ , namely the low OOM value of 2. The proper way of viewing Uniform(0,100) in our context is to think of it as a mixture consisting of 1% from the Uniform(0,1) plus 99% from the Uniform(1,100). The new measure of OMV for Uniform(0,100) yields the modest value of  $\text{LOG}[(90\text{th percentile})/(10\text{th percentile})] = [\text{LOG}(90) - \text{LOG}(10)] = [1.95 - 1.00] = 0.95$ .

Let us also compare the results of 'natural and uniform' as well as of 'odd and uniform' with a decisively non-uniform price list that exponentially explodes upwards, having numerous small prices, and few large ones. Here is one such example where each consecutive value is simply twice the previous one:

Random Linear Combination = List\*dice1 + List\*dice2

List (**geometric progression**) = {2, 4, 8, 16, 32, 64, 128, 256, 512}

LD = {32.7, 19.6, 13.0, 7.5, 9.2, 5.0, 5.8, 5.6, 1.7} **SSD = 28.8**

The near logarithmic result here (except for digit 9) can be naively thought of as due to the sufficiently large OOM of resultant data which is  $\text{LOG}(2*6*512/2*1*2)$  or 3.2, although if only OOM of the price list itself is measured, it yields  $\text{LOG}(512/2)$ , or 2.4. This explanation though is not quite correct, since  $2*6*512$  and  $2*1*2$ , as well as many other possibilities on the edges should be excluded from OOM calculations as outliers and such — in the spirit of the earlier qualification to the OOM principle. Applying OMV measure here is a bit involved, requiring that all possible purchases are displayed and sorted, 90th percentile and 10th percentile calculated, and data is purged of the two edges before LOG calculations. An alternative point of view of the near logarithmic result of the ‘geometric progression’ price list, is that it could be thought of as emanating from the close connection between geometric progression and Benford’s Law. In other words, that the digital distribution of resultant/generated data is also being influenced in general to some extent by the digital configuration of the price list itself, and its strong association with multiplication processes (with 2 being the multiplicative factor).

We have focused solely on first-order digits distribution in the study of Random Linear Combinations, yet it is worthwhile to explore second- and higher-order distributions, as well as last-two-digits combination. A case in point where such further exploration can be quite fruitful is when the price list has a particular predisposition to have certain repetition of multiples of integers, or other peculiarities (in other words, essentially when the price list is ‘too discrete’ to turn into something resembling continuous data under the transforming process of RLC). For example, consider a very peculiar retail shop where values in its 20-item price list are all multiples of five. In this simulation model we assume that the typical shopper buys at least two products, while tossing a coin to decide whether or not to purchase a third product. A ‘head’ reading on the coin induces buying and it is valued as 1, while a ‘tail’ reading inhibits buying and it is valued as 0.

Random Linear Combination = List\*dice1 + List\*dice2 + List\*dice3\*Coin  
 List (**multiples of five**) = {\$5, \$10, \$15, \$20, . . . , \$85, \$90, \$95, \$100}

Here, first-order digit distribution is quite different from the logarithmic, yet second-order distribution is quite close to the theoretical proportion. The most peculiar result here is the **last digit** distribution, supposed to be 10% for each digit according to Benford’s Law, but which shows extreme spikes for digits 0 and 5, leaving all other digits with zero proportions. Surely any product combinations

of 0 or 5 yields the closed set of 0 and 5! For example, if a shopper buys three of the item priced \$45, two of the item priced \$15, and one of the item priced \$95, the total bill is  $3*(45) + 2*(15) + 1*(95) = 135 + 30 + 95 = 260$ . Another shopper buying five of the item priced \$25, and four of the item priced \$20, will pay the total bill of  $5*(25) + 4*(20) = 125 + 80 = 205$ . The last digit of the bills of both shoppers is either 0 or 5. Here third-order distribution is almost identical to last-digit distribution since  $\approx 97\%$  of resultant data falls under \$999, while only  $\approx 3\%$  is over \$1,000. Third-order digits distribution here also shows extreme spikes of 59.7% and 38.0% for digits 0 and 5 respectively, leaving all other digits with near-zero proportions. Figure 3.28 shows the relevant first-, second-, and last-digit distributions for the above RLC scheme of ‘multiples of five’ price list.

The digital analyst examining revenue data here on the basis of first digits only might hastily conclude that the company publishes fraudulent accounting data. Yet, when higher orders as well as last-digits distributions are examined, they indicate that the company is probably honest.

The resiliency and stability of second-order digits (as compared with first-order digits) is a much more general principle, and the example of Fig. 3.28 is but one manifestation of this fact. This principle shall be further discussed in the theoretical fourth section. Beyond just a small shop selling in multiples of five dollars, second-order digits are almost always more steady, having by far less variability. For example, RLC model of {narrowly focused\*dice1 + narrowly focused\*dice2} seen earlier, with its strong first-digits deviation, is much more in accordance with

Digit	1st Order		2nd Order		Last Digit	
	Simulation	Benford	Simulation	Benford	Simulation	Benford
0			13.2%	12.0%	60.1%	10.0%
1	14.9%	30.1%	11.0%	11.4%	0.0%	10.0%
2	13.8%	17.6%	11.1%	10.9%	0.0%	10.0%
3	16.0%	12.5%	9.3%	10.4%	0.0%	10.0%
4	14.9%	9.7%	9.6%	10.0%	0.0%	10.0%
5	12.8%	7.9%	10.3%	9.7%	39.9%	10.0%
6	10.3%	6.7%	8.7%	9.3%	0.0%	10.0%
7	7.7%	5.8%	8.3%	9.0%	0.0%	10.0%
8	6.2%	5.1%	9.7%	8.8%	0.0%	10.0%
9	3.3%	4.6%	8.9%	8.5%	0.0%	10.0%

Figure 3.28 Digital Results of RLC When Price List Consists of Multiples of 5

Benford's Law of the second order, where distribution is {11.9, 11.0, 12.9, 10.3, 9.9, 10.7, 8.7, 8.8, 8.6, 7.3}. The RLC model {variable but pooled\*dice1 + variable but pooled\*dice2} which also deviates strongly in first-order sense, yields second digits of {11.3, 10.7, 14.7, 7.5, 6.2, 7.2, 10.3, 16.8, 10.3, 5.0}. Finally, RLC model {uniform(0,100)\*dice1+uniform(0,100)\*dice2+uniform(0,100)\*dice3} which gave decisively non-logarithmic first-digits distribution, nonetheless is much more compatible with the second, coming at {11.7, 10.7, 10.3, 9.9, 9.6, 9.9, 9.4, 9.2, 9.8, 9.5}. The three results above should be favorably compared with Benford's Law for the unconditional second order which is {**12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5**}.

Surely third and fourth orders are nearly equal and even more resilient than the second order, although in the context of fraud detection this typically cannot be utilized, since most fraudsters invent their fake numbers with digital equality in any case, either unintentionally in a totally random way which leads to equality by default, or intentionally mistakenly believing that digital equality exists in all real accounting data. Therefore a comparison between observed digital equality in fake accounting data and the theoretical equality of the law couldn't reveal any of the supposed fraud taking place.

[A note on the last digit: for data set with small numbers digit-wise, where almost all numbers are no more than three-digits wide, as in the above RLC case of multiples of 5, OOM is about 2, and this normally implies that it cannot be Benford in the first-order sense, but perhaps it is Benford in the second-order sense. In such cases, last-digit distribution should be superseded by third-digit distribution. In other words, observed distribution of third digit here is by definition also that of the last digit, yet last digits should be compared with Benford's third order, not with Benford's last order. Third-order Benford is ever so slightly different than the uniform 10% of the last-digit distribution. Last-digit distribution, by definition, is the one so far to the right, and of such high order that almost all digital skewness is gone and dissipated, and a near-perfect digital equality is achieved. Third-digits order is not sufficiently near that 10% exact digital equality].

We end this chapter with a generic discussion of what in essence leads to logarithmic convergence in revenue data sets by way of RLC processes. If a typical shopper goes out in the morning determined to buy one loaf of bread at the small local bakery, the price paid has very little variability, hence order of magnitude from highest to lowest possible bread expenditure is quite small. The shopper may

buy the bread at a price ranging from \$1 for a simple French baguette, to \$7 for fancy dark Russian bread. Still, the variability is low, and range of possibilities is narrow. On the other hand, for a shopper entering a very large store in general for an hour or more of extended shopping, not aiming solely at bread, the variability is huge, especially if several quantities of several distinct products may be purchased, that is, if it corresponds to our model of Random Linear Combinations.

Yet, a different digital picture emerges when the focused shopper in search of just one item anywhere within his or her large metropolis, is in need of purchasing not bread, but a very sophisticated and expensive item containing within itself a huge number of components, such as a car or a computer. Even though there is no variability at all in the type of product to be purchased nor in its quantity, as only one quantity of a car or a computer is to be purchased, the fact that it contains many costly components, each having small variability in pricing, all adds up to a large resultant variability. Hence a government census on millions of retail bread purchases countrywide should not obey Benford's Law, while the same census on millions of car or computer purchases should obey the law closely. Let us compare the two logarithmic purchasing events:

$$[\text{General Store Shopping}] = N_1 * \text{Item1} + N_2 * \text{Item2} + N_3 * \text{Item3}$$

$$[\text{Car}] = 4 * \text{Brake} + 1 * \text{Engine (with its many smaller components)} + 2 * \text{Bumper} \\ + 4 * \text{Door} + 1 * \text{Fuel Tank} + 3 * \text{Mirror} + 1 * \text{Transmission System} \\ + 4 * \text{Bearing} + 4 * \text{wheel} + 1 * \text{Air Conditioning} + 1 * \text{Steering System} \\ + 4 * \text{Tire} + 2 * \text{Air Bag} + \text{etc.}$$

Both are logarithmic under certain conditions, because both processes generate enough variability in resultant data. Both typically have an overall large order of magnitude. Store shopping involves no more than, say, three or five items, with quantities of purchased items varying only from, say, one to six, but its large variability springs from the huge number of items on sale, namely that long and varied price list at the large shop or supermarket. Each of the components in the car or computer has perhaps very little price variability, yet since they are so numerous, resultant variability of their sum is quite large, and hence the purchase is logarithmic.

## CASE STUDY X: FORENSIC ANALYSIS OF REVENUE DATA FOR SMALL SHOP

---

---

The model of Random Linear Combinations shall be applied in forensic digital analysis of reported revenue data for a very small shop with a short price list. The shop named Bohemian Crystal Palace sells glass crystals and other jewelries, and it is based in the Czech Republic. Its website is <http://www.crystal-treasury.com/index.html>. Hypothetical revenue amounts from its separate and specialized unit that sells bracelets are imagined to be reported to the Czech Tax Authority, containing hundreds of thousands of sales records covering a period of five years. The electronic database contains only the total amount due/paid for each invoice. More extensive information such as details about items and quantities purchased for each invoice, customer name, method of payment and so forth, are accessible only as paper-based information, and which is highly time consuming and costly to utilize. There are 13 distinct bracelets for sale in this separate company unit. In the main web page, the icon named “Enter Store” on the top left is selected, then “Jewelry” around the top right area is selected, followed by the selection of “Bracelets”. The Czech Tax Authority has assigned the data analyst with the task of determining whether or not financial fraud had been committed by the company, but has provided the analyst with very limited monetary resources for the job, and thus efficiency in the performance of the task is quite essential here. The distribution of the first digits of the electronic data is easily calculated by the data analyst and yields the proportions of {17.2%, 6.9%, 8.7%, 11.8%, 7.8%, 5.5%, 13.8%, 17.1%, 11.2%}. The data analyst wishes to use RLC model to obtain the expected (theoretical) digital configuration, given the available price list of the company, and which would then be compared to the empirically obtained digital results above. In order to apply RLC modeling, it is necessary to make assumptions on how many items and quantities are typically bought. This can be easily done without costly expenses or time consuming effort by simply taking a truly random sample of say 500 or so paper invoices to record typical occurrences of quantities and items chosen. Such study was undertaken, and the sample of invoices indicates that most

customers ( $\approx 70\%$ ) purchase only one item and choose anywhere between one and nine quantities equally; while a minority of them ( $\approx 30\%$ ) purchase two items and choose anywhere between one and five quantities for each item selected. Since the clientele is broken into two distinct groups, two distinct RLC models are performed with the weights of 0.7 and 0.3 respectively given to each in the aggregate.

Digits of reported revenue data: {17.2, 6.9, 8.7, 11.8, 7.8, 5.5, 13.8, 17.1, 11.2}.

The price list in U.S. Dollars is:

List (**Bohemian Crystal Palace**) = {166.11, 174.89, 178.85, 192.23, 239.16, 280.00, 280.00, 382.96, 596.81, 1266.24, 1283.10, 1655.41, 1703.04}

(I) Random Linear Combination = List\*Uniform{1, 2, 3, 4, 5, 6, 7, 8, 9}  
 LD = {43.3, 12.1, 10.3, 4.3, 8.5, 5.1, 5.2, 7.9, 3.4}

(II) Random Linear Combination = List\*Uniform 1 {1,2,3,4,5} +  
 List\*Uniform 2 {1,2,3,4,5}  
 LD = {32.1, 15.3, 9.4, 7.4, 9.2, 7.1, 7.7, 6.2, 5.6}

The following calculation yields the aggregate expectation of first-digit configuration:

$$\begin{aligned}
 &0.7 * \{43.3, 12.1, 10.3, 4.3, 8.5, 5.1, 5.2, 7.9, 3.4\} + \\
 &0.3 * \{32.1, 15.3, 9.4, 7.4, 9.2, 7.1, 7.7, 6.2, 5.6\} \\
 &\text{-----} \\
 &\{39.9, 13.1, 10.0, 5.2, 8.7, 5.7, 5.9, 7.4, 4.1\}
 \end{aligned}$$

This theoretical result is certainly not compatible with the empirical result of provided revenue data, and the analyst has informed the Czech Tax Authority of the strong possibility that fraud has been committed. Upon learning of the damning conclusion arrived by the analyst, the company sent a memo, claiming that in about 95% of purchases only one item is bought, with no more than three quantities being selected; that single purchases of nine, for example, or double purchases of five quantities are extremely rare occurrences; and that therefore the method of digital analysis has been flawed, applying the wrong assumptions. This had the analyst worried that the company might indeed be honest, and that perhaps by some rare chance the sample of 500 invoices was misleading and in error. The analyst then set out to examine whether applying the company's claim of purchases of single item with quantities chosen from the Discrete Uniform on {1,2,3} would somehow

match the observed digital configuration. The following four separate RLC simulation results, with 7,000 runs each, came out as follows:

**(III)** Random Linear Combination = List\*Uniform {1,2,3}

LD = {28.6, 12.5, 23.0, 7.6, 17.9, 0.1, 5.4, 4.9, 0.0}

LD = {28.1, 12.6, 23.8, 7.6, 17.8, 0.2, 5.0, 4.9, 0.1}

LD = {28.9, 12.3, 22.9, 7.7, 17.8, 0.1, 5.2, 5.0, 0.1}

LD = {28.2, 12.6, 22.6, 8.0, 17.8, 0.1, 5.5, 5.1, 0.0}

These results regarding first-order distribution strongly refuted the new claim made by the company. These four simulation results also showed a peculiar yet very consistent digital signature for the RLC model. The existence of a consistent digital configuration or signature is a universal property for all data sets, processes, and distributions, regardless whether they are logarithmic or non-logarithmic, random or deterministic.

**Section 4**

**CONCEPTUAL  
AND MATHEMATICAL FOUNDATIONS**

**This page intentionally left blank**

## HYBRID DATA SETS BLENDING SEVERAL DATA TYPES

---

---

The theoretical and conceptual investigation into the whole digital phenomenon of Benford's Law starts with the examination of a special class of random data types derived by the aggregation, compilation and mixing of several distinct data sets. Such mixing and compilation of a variety of (related or totally unrelated) data sets, leads to a strong logarithmic convergence in and of itself under certain conditions, and serves as a prominent cause for the prevalence of Benford's Law in real-life data sets.

As an example, we consider data on infectious disease cases per province or per city from all over the world, supplied to the World Health Organization by all member countries. When the assigned statistician at the WHO headquarters organizes all input information on Hepatitis C virus, for example, into one large computer file as one single set of data, he or she is creating just such aggregation of data sets, merging the cases of 194 different countries into a single global case. The statistician typically separates international data into categories by type of bacteria or virus involved, or might be curious to put all types of pathogens into one large file and analyze results, thus increasing the degree of aggregation by one notch. An even better example is given by the overly enthusiastic sport statistician who attempts to create some kind of pan-sport database by considering results of a particular measure such as length of time of actual games played in football, tennis, baseball, basketball, and other sports, compiling it into one large computer file. If a variety of totally different sport measurements are then also aggregated, such as scores, length of time of games, ratings, and more, and are all put into one long column within one single Excel file, for example, it constitutes an even higher level of aggregation. An extreme form of aggregation example can be given perhaps by that errant and eccentric statistician who for some reason or another randomly compiles numbers from all the pages in all major newspapers, mixing totally unrelated numbers, quantities, and values, and goes on to include journals on any topic, as well as random pick from web pages on the Internet. That errant statistician actually turned out to be a very astute and capable digital analyst, well-trained in Benford's Law and all its aspects, since such collection of numbers obeys Benford's Law in the limit as its size grows to be infinitely large. Surely, the ultimate form of extreme aggregation is Aggregate Global Data, which indeed happened to form the motivation behind such interpretation of Benford's Law.

## SECOND-GENERATION DISTRIBUTIONS

---

---

Second-Generation Distributions (SGD) refers to the resultant distribution of any collection of, combination of, construction out of, standard statistical distributions. For example, a game may be designed to pick a single value from the distributions Uniform(0, 87), Normal(53, 2), and Exponential(64) by way of a throw of a dice; assigning the Uniform if 1, 2, 3, or 4 are thrown; the Normal if 5 is thrown; and the Exponential if 6 is thrown. Subsequent simulation from the chosen distribution finally determines the ultimate number in the game. Such a game ultimately gives birth to a unique distribution — existing in its own right — being built out of these three distributions who have fathered it with the aid of a dice, hence the name ‘second generation’. Surely the curve of the Probability Density Function (PDF) of the value in such a game would not look anything like one of its components, but rather its shape as well as its range should typically appear quite odd and very particular, heavily dependent (among other things) on the exact values of the parameters of the three distributions it is based on. It must be noted how one distribution (the dice) was used in pointing to one of the three other distributions, creating this statistical dependency. In another example, a set of four Normal distributions for human heights in units of meter pertaining to four different nationalities, Normal(1.85, 0.3), Normal(1.71, 0.2), Normal(1.93, 0.4), and Normal(1.59, 0.2) are to be considered equally likely, and a draw of a single value from one of these four distributions takes place. The process may be viewed as utilizing a throw of an imaginary unbiased dice having four sides to decide on the distribution. Clearly, some kind of simple averaging of the heights of the four density curves should yield the desired resultant density curve of the whole process. A classic case in point is the Central Limit Theorem, regarding the sum of independent and identically distributed random variables  $X_i$ , which is Normal in the limit as numerous such variables are incorporated via addition. Hence  $SUM = X_1 + X_2 + X_3 + X_4 + \dots$  metamorphosed into that very familiar Normal; but this is rare and typically more intricate or peculiar distributions are the results in other constructions and combinations of random variables. Instead of adding as in the Central Limit Theorem, one

may be multiplying, dividing, taking one variable to the power of another variable, and so forth. Yet, one of the most typical and important second-generation distribution in applications is the simple **average** (giving equal weights) of a variety of  $N$  distributions, which is indeed nothing but the **sum** divided by the (irrelevant) constant  $N$ , and such a scheme leads us directly to the Central Limit Theorem (except that — strictly speaking — all distributions should be identical).

We have indeed encountered already (unknowingly) some second-generation distributions when dealing with the Random Linear Combinations model applied specifically to company accounting revenue data, although in all RLC models the structure is quite rigid and limited in style. It must be noted that the number of possibilities for intricate setups of probability densities is infinite and third-generation distributions or higher could also be considered.

Second-generation distributions are quite frequently (though not always) logarithmic. Most crucially, many SGD that are not logarithmic in a strict sense are approximately so, and this constitutes yet another crucial source of the prevalence of the logarithmic distribution in numerous cases of real-life data sets.

## A LEADING DIGITS PARABLE

---

---

Imagine farmers and shepherds in ancient Greece around the year 1500 BC. They live very simple lives, and they have very few possessions. Their severe poverty implies that they possess not more than eight quantities per item — while the number 9 is used only with regards to the Gods, believing in 9 deities at that epoch. They utilize only very simple numbers, the ‘earthly numbers’ from 1 to 8 and the ‘godly number’ 9. They are ignorant of 0, 10, decimals, fractions; and most importantly, they are completely ignorant of square roots, true bliss!

Their mental lives and vocabulary are quite limited, they have only nine objects: Gods, spouse, sheep, houses, olives, chickens, dogs, oranges, and slaves. There was never any need to invent numbers beyond 9 as they never possess or contemplate anything more than 9. As it happened, there are various (self-imposed or natural) limits on the possession or existence of these nine objects. For example, believing in a 10th God is considered serious heresy punishable by death. Polygamy is another taboo hence 1 is the only number used in conversation regarding the topic of a wife or a husband. Due to a newly installed government with humanist tendencies, new laws were enacted limiting slave exploitation to 3. Other limits on their possessions are: 8 olives, 7 dogs, 6 chickens, 5 sheep, 4 oranges, 2 houses.

Hence the backdrop is set for digital analysis of this particular society that coincidentally utilizes exactly only **9 numbers** as well as referring exactly to only **9 topics**. For them numbers and digits are interchangeable, since digit 7 ‘leads’ number 7, and so forth. Digital leadership for them depends on the specific topic of their simple conversations. For example, for conversations about spouse, they use only the number 1, and for that topic digit 1 leads 100% of the time, while all other digits lead by 0%. For slaves, being that no one is allowed to possess more than three; digits 1 to 3 lead by 1/3 each, while digits 4 to 9 do not lead at all. For Gods, all digits lead equally by 1/9, namely by 11%, and the more progressive elements in their society point to this fact as a divine sign that ethnic and racial equality should rule the land.

Here are some typical conversations:

I have only 1 wife, and I prefer it that way.

I will give you 3 of my sheep tomorrow if you help me today.

I have lost 4 oranges yesterday near Athens.

He has got 7 Gods on his side while I have got only 2 on my side (Zeus & Venus), it's useless to oppose him.

Yesterday I saw 5 dogs chasing 3 chickens.

Let us attempt to aggregate all types of conversations and examine overall resultant digit distribution, in the spirit of Aggregate Global Data Interpretation of Benford's Law. It is necessary to make an assumption about the relative importance (frequency) of each topic. Since topic equality is deemed reasonable, namely that they talk about spouse, slaves, Gods, and other topics in equal proportions, then it shall be assumed. It is also necessary to make an assumption about the relative occurrences (frequency) of quantities within each topic. Are they more likely to talk about 3 olives than say about 8 olives? Let us assume that all quantities within a topic come with equal probability. Under these assumptions, the model is complete, and overall digit configuration can be calculated by simply taking the average digital proportions of all the topics digit by digit. These calculations are shown in Fig. 4.1(A).

The explanation for the digital disparity can be summed up by the simple comparison of digit 1 with digit 9; digit 1 turned out to be a very popular digit naturally as it occurs in all topics; digit 9, on the other hand, occurs only in religious conversations.

As a general check on the above result, a more convincing approach perhaps is to simply simulate nine typical conversations for each topic, and then calculate

Digit	Spouse	Houses	Slaves	Oranges	Sheep	Chicken	Dogs	Olives	Gods	Averages
1	1/1	1/2	1/3	1/4	1/5	1/6	1/7	1/8	1/9	----> 31.4%
2	0	1/2	1/3	1/4	1/5	1/6	1/7	1/8	1/9	----> 20.3%
3	0	0	1/3	1/4	1/5	1/6	1/7	1/8	1/9	----> 14.8%
4	0	0	0	1/4	1/5	1/6	1/7	1/8	1/9	----> 11.1%
5	0	0	0	0	1/5	1/6	1/7	1/8	1/9	----> 8.3%
6	0	0	0	0	0	1/6	1/7	1/8	1/9	----> 6.1%
7	0	0	0	0	0	0	1/7	1/8	1/9	----> 4.2%
8	0	0	0	0	0	0	0	1/8	1/9	----> 2.6%
9	0	0	0	0	0	0	0	0	1/9	----> 1.2%

Figure 4.1(A) Digit Distribution in Ancient Greece

aggregate digital proportions for all 81 conversations. Here is one such attempt (artificially beautified and manipulated a bit for pedagogical purposes):

Spouse {1, 1, 1, 1, 1, 1, 1, 1, 1}  
 Houses {1, 2, 1, 2, 1, 2, 1, 2, 1}  
 Slaves {1, 2, 3, 1, 2, 3, 1, 2, 3}  
 Oranges {1, 2, 3, 4, 1, 2, 3, 4, 4}  
 Sheep {1, 2, 3, 4, 5, 1, 2, 3, 5}  
 Chicken {1, 2, 3, 4, 5, 6, 3, 5, 6}  
 Dogs {1, 2, 3, 4, 5, 6, 7, 7, 4}  
 Olives {1, 2, 3, 4, 5, 6, 7, 8, 2}  
 Gods {1, 2, 3, 4, 5, 6, 7, 8, 9}

Counting overall digit occurrences we get {25, 16, 12, 9, 7, 5, 4, 2, 1}. Proportions are {25/81, 16/81, 12/81, 9/81, 7/81, 5/81, 4/81, 2/81, 1/81}, namely {**30.9%**, **19.8%**, **14.8%**, **11.1%**, **8.6%**, **6.2%**, **4.9%**, **2.5%**, **1.2%**}. This result is very similar to the result in Fig. 4.1(A).

But surely, the parable told here constitutes a type of hybrid data blending several data sets. It also corresponds to a second-generation distribution, as each topic points to a particular uniform discrete distribution. For example, sheep has a discrete uniform distribution of the set {1, 2, 3, 4, 5} with 1/5 equal probability for each quantity. The topic of slaves has a discrete uniform distribution of the set {1, 2, 3} with 1/3 equal probability for each quantity. Since we have assumed equality in occurrences of topics, a fair imaginary dice of nine sides must be constructed, and then the simulation game can begin in earnest.

Taking stock: the relevant model here is a collection of nine distinct uniform discrete distributions where each (topic) is accorded equal importance. All distributions start at 1 being the **Lower Bound (LB)** for all of them, hence 1 is the anchor orienting, synchronizing and uniting all 9 distributions in a certain position. Length of distributions, that is, their **Upper Bound (UB)** is gradually being increased from 1, in the spouse case where  $LB = UB = 1$ , all the way to 9, in the Gods case where  $LB = 1$  and  $UB = 9$ .

Could this model be relevant at all to Benford's Law in our modern life? Surprisingly, the answer is a resounding 'yes', albeit with some reservations, qualifications, and modifications! It is doubtful that it could serve as an overall highly successful model in terms of the Aggregate Global Data Interpretation of Benford's Law for the modern era. Nonetheless, it does serve as a good model for numerous modern data sets of a particular type, and in spite of the fact that it is a quite

restricted model, with 9 as the UB for all distributions, as opposed to much higher ranges of values in our modern everyday data, with an upper limit in the millions, billions, or trillions. The Greek model can be applicable in spite of the fact that it does not incorporate fractions, decimals, ratios, negative numbers, or the greatly feared and avoided square roots. Remarkably, with a few modifications and improvements, the Greek Parable leads directly (arithmetically, that is) to the logarithmic distribution as shall be demonstrated in the next few chapters, and with the rationale behind it as an appropriate model representing at least some particular types of real-life data still intact and valid. To recap: at a minimum, a good part of typical real-life modern data can be represented by a model built on a large collection of discrete uniform distributions with varying upper bound, all sharing that common anchor 1 as lower bound, serving to orient numbers in much the same way as for those ancient Greeks in the parable above.

Skeptics, purists, and doubters might contend that the underlying assumptions of such society have been artificially and deliberately manipulated so as to arrive at something quite close to Benford's Law. Such criticism of the Greek Parable proposes an **inverted** set of limitations on object availability and societal rules as follows:

(I) A fanatical belief in a 9-deity-combination as God. The mere mentioning of, say, 5 Gods or monotheistic single God, is severely punished. All religiously-correct conversations involve digit 9 exclusively. (II) Marrying up to 9 spouses (wives or husbands) is allowed and even strongly encouraged. Thus typical marital occurrences are equally distributed on  $\{1, 2, 3, \dots, 9\}$ . (III) Newly enacted "*House Hospitality Law*" forbidding the ownership of a single house as it symbolizes selfishness and lack of extra space for potential guests. This societal rule is in addition to the natural limit of 9 houses per family. Therefore house ownership is equally distributed on  $\{2, 3, 4, \dots, 9\}$ . (IV) The slogan "*Slavery with Humanity*" reflects a long tradition in Greece of gentle and humane treatment of slaves, leading finally to a newly enacted law prohibiting the ownership of a single lonely slave or even a pair of slaves for any single household. It was thought that at a minimum a group of 3 or more slaves is needed to easily bear the harsh condition and difficult labor by socialization and the sharing of the hardship. Hence slave ownership is equally distributed on  $\{3, 4, 5, \dots, 9\}$ . Clearly one is running out of stories and imagination to explain why nobody is eating or allowed to eat, say, 5, 7, or 3 olives for breakfast, instead of the allowed consumption of either 8 or 9 olives exclusively. In any case, under these assumptions the resultant average digit distribution for such bizarre society is depicted in Fig. 4.1 (B).

Digit	Spouse	Houses	Slaves	Oranges	Sheep	Chicken	Dogs	Olives	Gods	Averages
1	1/9	0	0	0	0	0	0	0	0	-----> 1.2%
2	1/9	1/8	0	0	0	0	0	0	0	-----> 2.6%
3	1/9	1/8	1/7	0	0	0	0	0	0	-----> 4.2%
4	1/9	1/8	1/7	1/6	0	0	0	0	0	-----> 6.1%
5	1/9	1/8	1/7	1/6	1/5	0	0	0	0	-----> 8.3%
6	1/9	1/8	1/7	1/6	1/5	1/4	0	0	0	-----> 11.1%
7	1/9	1/8	1/7	1/6	1/5	1/4	1/3	0	0	-----> 14.8%
8	1/9	1/8	1/7	1/6	1/5	1/4	1/3	1/2	0	-----> 20.3%
9	1/9	1/8	1/7	1/6	1/5	1/4	1/3	1/2	1/1	-----> 31.4%

FIGURE 4.1 (B) Erroneous Digit Distribution — Artificially Inverted Assumptions

Or as a data set approximately:

- Spouse {1, 2, 3, 4, 5, 6, 7, 8, 9}
- Houses {6, 2, 3, 4, 5, 6, 7, 8, 9}
- Slaves {4, 8, 3, 4, 5, 6, 7, 8, 9}
- Oranges {5, 8, 6, 4, 5, 6, 7, 8, 9}
- Sheep {6, 9, 5, 8, 5, 6, 7, 8, 9}
- Chicken {9, 8, 6, 7, 8, 6, 7, 8, 9}
- Dogs {7, 8, 9, 7, 8, 9, 7, 8, 9}
- Olives {8, 9, 8, 9, 8, 9, 8, 9, 8}
- Gods {9, 9, 9, 9, 9, 9, 9, 9, 9}

Counting overall digit occurrences we get {1, 2, 3, 5, 7, 10, 10, 18, 25}. Proportions are {1/81, 2/81, 3/81, 5/81, 7/81, 10/81, 10/81, 18/81, 25/81}, namely {1.2%, 2.5%, 3.7%, 6.2%, 8.6%, 12.3%, 12.3%, 22.2%, 30.9%}. This result is very similar to the result of Fig. 4.1(B).

Such resultant digit distribution is [extremely] contrary to Benford’s Law. Yet, the original setup of the Greek Parable as shown in Figure 4.1 (A) is the **most natural** and the **most typical** in real-life data sets. It is the latter alternative setup of Fig. 4.1 (B) that is artificially inverted, and even though it might apply in some very particular situations, it is very rarely found.

Another attempt in sabotaging the parable is the proposition of equality on societal and physical limitations for all objects; 9 being that universal limit, which leads to the table in Figure 4.1(C).

Digit	Spouse	Houses	Slaves	Oranges	Sheep	Chicken	Dogs	Olives	Gods	Averages
1	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	-----> 11.1%
2	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	-----> 11.1%
3	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	-----> 11.1%
4	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	-----> 11.1%
5	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	-----> 11.1%
6	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	-----> 11.1%
7	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	-----> 11.1%
8	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	-----> 11.1%
9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	-----> 11.1%

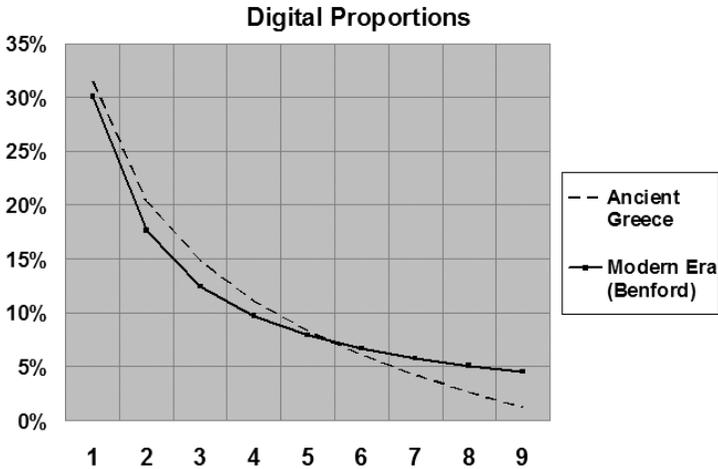
Figure 4.1 (C) Another Erroneous Digit Distribution — Equality of Limits Assumption

Or as a data set:

- Spouse {1, 2, 3, 4, 5, 6, 7, 8, 9}
- Houses {1, 2, 3, 4, 5, 6, 7, 8, 9}
- Slaves {1, 2, 3, 4, 5, 6, 7, 8, 9}
- Oranges {1, 2, 3, 4, 5, 6, 7, 8, 9}
- Sheep {1, 2, 3, 4, 5, 6, 7, 8, 9}
- Chicken {1, 2, 3, 4, 5, 6, 7, 8, 9}
- Dogs {1, 2, 3, 4, 5, 6, 7, 8, 9}
- Olives {1, 2, 3, 4, 5, 6, 7, 8, 9}
- Gods {1, 2, 3, 4, 5, 6, 7, 8, 9}

Surely such setup is just as artificial and far removed from typical everyday data sets as the inverted setup of Fig. 4.1 (B). Each topic, object, variable, and measurement typically has its own distinct limit and natural range; and proposing a common spread across diverse data types is extremely unnatural and rare.

No elaborate mathematical proof is presented here, nor any Monte Carlo computer simulation. Absent also is any scientific laboratory experiment or empirical data analysis demonstrating that only the original setup of Figure 4.1 (A) is the proper and the most natural one. Rather the reader is asked to exercise his or her intuition, common sense judgment, and very general knowledge or experience with real-life data in order to arrive at the correct conclusion; namely that most Lower Bounds typically congregate together around 0, 1 or some very low value, while Upper Bounds vary wildly and are quite different for each topic or variable.



**Figure 4.2** Digit Distributions — Ancient Greece vs. Modern Era

Figure 4.2 demonstrates the overall similarity and dissimilarity between the result of the parable and the logarithmic.

Finally, it is important to observe the universal nature of the applicability of the Greek parable, as seen in its total independence of any scale or units of measurements, such as the kilogram, milligram, pound, meter, or feet. Moreover, the parable is also totally independent of any choice of a base within the number system, such as 10, 16,  $e$ , or 2. Since nothing was assumed regarding choices for scale or base, the parable represents a scale-invariant as well as base-invariant argument in favor of a digital distribution that is fairly close to the logarithmic.

## SIMPLE AVERAGING SCHEME AS A MODEL FOR TYPICAL DATA

---

---

Our civilization at the current epoch is certainly much more sophisticated than the society described in the Greek parable. We have use of vastly more topics of conversations than that meager nine-item set, and we have greatly expanded our number system well beyond their limit of nine whole numbers. We have also advanced beyond their simple count data and now use units and scales to record weights, lengths, time, and so forth. Therefore, if we were to attempt then to expand on the Greek parable to include the immense variety of typical modern speeches and recorded data, we would obviously face a daunting or rather impossible task. How does one go about accounting for and aggregating digital configurations of, say, stock prices, river lengths, accounting expense data, weights of people, accidents per year, and so forth, to mention only a tiny portion of everyday data? Could an argument be made purporting to show that a modified model of the Greek parable corresponds well to Aggregate Global Data in the context of such an interpretation of Benford's Law? Real-life modern data sets often do not even start near the value of 1, and therefore that unifying lower bound anchor of 1 for all topics in the Greek parable is frequently absent here, as numerons data sets start at a higher minimum level. Also, upper bounds vary considerably and wildly as opposed to simply increasing nicely one integer at a time as in the parable. Moreover, the parable's two assumptions of probabilistic equality of values within each Uniform distribution, as well as equality of importance between them, are not appropriate here. It may be argued though that non-equality (within and between) possibly cancels out for such vast quantity of data, that there exists some grand trade-off here preventing such inequality from favoring low digits over high digits or vice versa as it works both ways. It may also be argued that not starting exactly near 1 is immaterial in the grand scheme of things; that real-life data sets tend to start near very low values and to end gradually around much higher values (all of which resembles the parable); but such arguments are a bit vague and uncertain. Nonetheless, let us extrapolate from the Greek parable, and develop a limited but well-structured mathematical

model that is capable of capturing or representing at least an important class of usage of numbers in the modern world.

And so we consider what would happen collectively to leading-digits distributions of numerous intervals, all starting at 1, made only of the integers, while differing in their length. Those intervals should be made progressively longer by systematically increasing their length exactly by one integer at a time. The plan is then to obtain an aggregate digital distribution representing *all* the intervals, by simply taking the (unweighted) average of the digital distributions of all the intervals. For example, in the Greek parable there are nine such intervals:  $\{1\}$ ,  $\{1,2\}$ ,  $\{1,2,3\}$ ,  $\{1,2,3,4\}$ ,  $\{1,2,3,4,5\}$ ,  $\{1,2,3,4,5,6\}$ ,  $\{1,2,3,4,5,6,7\}$ ,  $\{1,2,3,4,5,6,7,8\}$ , and  $\{1,2,3,4,5,6,7,8,9\}$ . It should be noted that we are **not** calculating directly digital distribution of all the integers here mixed together as in  $\{1,1,2,1,2,3,1,2,3,4,1,2,3,4,5,1,2,3,4,5,6,1,2,3,4,5,6,7,1,2,3,4,5,6,7,8,1,2,3,4,5,6,7,8,9\}$ . Rather, we are averaging out 9 different digital proportions for those nine intervals, a totally different concept, yielding very different numerical result. The Greek parable averages out all nine distributions  $\{1,0,0,0,0,0,0,0\}$ ,  $\{0.5, 0.5, 0,0,0,0,0,0\}$ ,  $\{0.33, 0.33, 0.33, 0,0,0,0,0\}$ ,  $\dots$ ,  $\{0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11\}$ .

The individual elements of this model are then the sets of consecutive integers (intervals) on  $[1, \text{UB}_i]$  all accorded equal probability of occurring within the entire model itself. This last requirement should be noted carefully since it implies that digital distribution of the shortest interval, say,  $[1, 5]$  gives as much weight to the overall distribution as that of the longest interval, say,  $[1, 999]$ , and in spite of the fact that  $[1, 5]$  contains by far fewer integers. It is now then necessary to decide on the starting and ending values of  $\text{UB}_i$  (the upper bounds); in other words, we need to decide on the size of the smallest interval and on the size of the largest interval. In the Greek parable,  $\text{UB}_i$  varied only from 1 to 9, and it surely needs expanding. In order to decide on this issue, let us first explore the ramification of having a particular  $\text{UB}$  as the highest integer for a *single* individual interval, and in particular let's compare the two intervals  $[1, 3000]$  and  $[1, 9999]$ . For the former case where  $\text{UB}$  equals 3000, digits 1 and 2 have strong advantages and each leads 1112 numbers out of 3000 total, or  $1112/3000$ , that is 37% each, 74% for 1 and 2 combined, while all other digits lead with probability of  $111/3000$ , merely 3.7%, a biased situation we may wish to avoid. On the other hand, for the latter case where  $\text{UB}$  equals 9999, and indeed for all cases with  $\text{UB}$  as a number composed entirely of nines, such as 99, 999, 9999 and so forth, all nine digits have equal  $1/9$  probability of leading, an equitable situation that is quite preferable since it does not bias the

model towards any particular digit in the first place. A moment's thought would then convince everyone that letting the UB of the *largest* interval in the entire scheme fall far from some number made of nines would bias the model towards low digits unfairly. Equally unfair is to let the UB of the *smallest* interval fall far away from some number made up of nines. In conclusion: UB<sub>i</sub> should vary from a number of nines to another (higher) such number, or else from 1 itself (the LB!) to some large enough number of nines, or as should be quite obvious by now, to an IPOT number such as 10, 100, 1000, and so on. We are not arguing that real-life data itself tend to abruptly terminate exactly at 100, 1000, or other IPOT numbers, this is obviously not the case; instead we are merely trying to calibrate our mathematical model so as to avoid any bias or arbitrariness as much as possible.

It should be noted that the exclusive use of integers does not limit the relevance of the model to real-life data. Fractions can be easily incorporated and attached to each integer in the model without altering in the least the arithmetical results of first-order distribution. In addition, the model manages to imitate real-life data in the sense that it comes with this large **variability in UB** as compared with the relative **stability of LB** which typically in real data is stuck at 0 or 1 or just some slightly higher number.

The model described above shall be called '**simple averaging scheme**'. Figure 4.3 is the table of computerized results of a few such schemes, all with LB stuck at 1. Results here only resemble the logarithmic, but they are not quite close enough. The Greek parable is shown in the table on the left as '1-9'. A definite distribution limit exists for the scheme, and it is quickly reached as ranges of UB

Ranges of Upper Bound:	1 - 9	50 - 700	10 - 100	1 - 100	1 - 10,000	1 - 100,000
Digit 1	31.4	28.6	24.5	25.1	24.2	24.1
Digit 2	20.3	20.5	18.4	18.6	18.3	18.3
Digit 3	14.8	15.2	14.5	14.6	14.5	14.5
Digit 4	11.1	11.3	11.7	11.6	11.7	11.7
Digit 5	8.3	8.1	9.4	9.3	9.5	9.5
Digit 6	6.1	5.3	7.6	7.4	7.6	7.6
Digit 7	4.2	3.9	6.0	5.8	6.0	6.0
Digit 8	2.6	3.7	4.6	4.4	4.6	4.7
Digit 9	1.2	3.5	3.3	3.1	3.4	3.4

Figure 4.3 Simple Averaging Schemes for Various UB layouts with LB at 1 (CITY)

increase towards infinity. In the limit, where ranges of UB are quite large while spanning IPOT values, this distribution is known as **Stigler's Law**, shown in the last column on the right with precision of just one decimal place. It is named after George Stigler who argued against Benford's proof, assumptions, and conclusion, and developed his own mathematical model for the first digits. His proposed alternative first-digit distribution is expressed as  $[d*\ln(d) - (d + 1)*\ln(d + 1) + 3.55843]/9$ , with 'ln' denoting the natural e based log, and which is in perfect agreement with our simple averaging scheme in the limit. **Stigler** claimed that the correct first-leading-digits distribution should be given instead by  $\{24.1\%, 18.3\%, 14.5\%, 11.7\%, 9.5\%, 7.6\%, 6.0\%, 4.7\%, 3.4\%\}$ . The biased scheme with UB range of 50 to 700 was intentionally included here, as an illustrative odd case, where results are actually not too different in spite of violating that IPOT rule!

Hence the model in its present form **cannot** explain Aggregate Global Data and its related interpretation of Benford's Law. However, to explain the discrepancy between this result (Stigler's) and the logarithmic, an argument can be raised that as complex as the scheme seems to be, it is still not complex enough compared with the totality of everyday data. In other words, that it's still too narrow and specific corresponding solely to the particular chosen setup of lower and upper bounds. In a nutshell: it doesn't average enough things, hence its resultant digital distribution lacks universality.

For a better understanding of what is lacking in the averaging model above, let us examine address data, referring specifically to the house number. As seen earlier, a random sample from the Yellow Pages in South Dakota strongly points to a tendency towards the logarithmic. House number on all streets starts at 1, but street length varies. Some streets are short, some are long, and therefore the house with the highest numerical value is typically different for each street. Hence only UB is to vary, not LB which has to be fixed at 1. Clearly, our model of the simple averaging scheme would be a perfect representation of address data if we knew the length of the shortest street and that of the longest street, and also if we were assured that all street lengths in between those two extremes are distributed uniformly, namely a steady increase in length. Here, each physical street corresponds to an interval  $[1, UB_i]$  in our abstract model. Yet, an important consideration here is that typically upper bounds for street addresses do not vary between IPOT numbers. Say, we analyze street data for the city of Chicago where street length varies smoothly and gradually from 6 to 34,098. The simple averaging scheme performed for the city of Chicago should be specially tailor-made for its street configuration, and thus UB's should start at 6 and terminate at 34,098, not between any IPOT

numbers. Street length for the small town of Spring Valley in New York State may vary from 8 to 285, which is quite different than the street configuration of Chicago, and thus it is in need of a different averaging scheme. Of all the places in the USA, only in one rare and remarkable case of Chester, a small town in Pennsylvania, did the census office find street lengths varying exactly between 10 and 100. For Chester, a scheme running between these IPOT numbers is necessary, but for other cities UB must be running according to their own specific street configuration. It is only for the generic simple averaging model that UB needs varying between IPOT numbers so as not to bias the scheme towards any particular digits.

But what if the U.S. Census office wishes to examine aggregate house numbers from all the cities and all the towns in the USA? In other words, to obtain country-wide data on street addresses. Since each city and town has its own unique street configuration, the statistician needs to account for the variety in street length. **Clearly, for an entire country, we need to average results from multiple simple averaging schemes having different UB layouts.** Hence one can adapt the following point of view: the simple averaging scheme was good enough for just one city, and the need for a more complex averaging scheme is essentially the need to get the aggregate digital distribution of a whole country for all its cities and towns. We should now assume uniformity of variation in street configuration for the set of all these cities and towns. Constructing that more complex averaging scheme would be the topic of the next chapter. The street address interpretation of the simple averaging scheme clearly explains the falsehood of the scenarios in Figs. 4.1(B) and 4.1(C). The former signifies an extremely strange city. The latter signifies an odd city where all streets are of the same length.

We end this chapter by noting the remarkable robustness of the generic simple averaging scheme albeit in a somewhat weak sense. Had we averaged with UB varying between 10 and 90 (instead of the more proper way between, say, 10 and 100) low digits would win even more as digit 9 would be further diminished here. Had we been overshooting 100 by setting UB varying between 10 and 110, low digits would still win, as digit 1 obtains an extra advantage on the interval (100, 110). Hence the general trend holds favoring low digits over high ones, no matter where UB terminates! This feature gives a great deal of stability and consistency to the simple averaging scheme. Finally, Stigler's Law should not be viewed as a failure or an error. The distribution is perfectly proper for address data of a large metropolitan where street length just happened to vary between IPOT values. The simple averaging model and its associated Stigler's Law has applications well beyond city address data, owing to the fact that there are other data types that can be modeled in the same way.

## MORE COMPLEX AVERAGING SCHEMES

---

---

The next level of complexity is averaging out multiple simple averaging schemes themselves, while gradually varying their focus. This is typically done by letting LB stay fixed at 1, letting UB start also at a fixed point for all schemes (min UB) while allowing UB to terminate on a variety of different locations (max UB). If applied to street address, then the implicit assumption here is that the shortest street (min UB) is the same for all cities and therefore fixed, although arithmetically this restriction doesn't affect results very much. In addition, when applied to street address, min UB naturally should be fixed at 1 or just slightly above 1 (being the shortest street), while max UB should be made to vary having much larger values as it stands for the limit of the longest streets, (the table in Fig. 4.4 violates this requirement in one case as it pertains to generic schemes). For example, consider multiple simple averaging schemes where LB is fixed at 1, min UB fixed at 10, while max UB varies from 100 to 1000. In other words, the simple schemes with UB: **10 to 100**, **10 to 101**, **10 to 102**, ..., **10 to 999**, **10 to 1000**, and which are then followed by the averaging of all these 901 different averages. This particular example is shown on the third column from the left of Fig. 4.4, along with few other such averages of averaging schemes.

Resultant digit distributions here are much superior to the simple averaging scheme in the sense that they are much closer to the logarithmic. Also of note here is the last column on the right, where the arbitrary numbers of 4500, 800, and 30 did not disrupt overall results by much in spite of the theoretical requirement that ranges should start and end on IPOT numbers to avoid any bias towards low or high digits.

We have successfully aggregated and measured Continental USA. It can be claimed then that this country-wide data represents something 'more random' in a sense than data for just one single city, and this may explain its better digital comparison with the logarithmic. It is possible that averaging of averages simply has higher order of magnitude than in the simple averaging case, which explains its superior digital configuration, as discussed earlier. Obviously, what we now seek is

max UB (Highest)	100	1000	1000	1000	100	4500
max UB (Lowest)	20	1	100	500	1	800
min UB	10	1	10	15	1	30
LB	1	1	1	1	1	1
<hr/>						
Digit 1	33.9	30.9	30.4	28.3	34.7	32.0
Digit 2	19.6	19.0	18.9	20.2	19.7	16.8
Digit 3	12.6	13.2	13.1	14.9	13.0	11.1
Digit 4	8.9	9.8	9.8	10.9	9.3	8.9
Digit 5	6.7	7.6	7.6	7.9	6.9	7.8
Digit 6	5.4	6.1	6.2	5.9	5.4	6.9
Digit 7	4.6	5.1	5.2	4.6	4.3	6.1
Digit 8	4.2	4.4	4.6	3.9	3.6	5.5
Digit 9	4.0	4.0	4.1	3.4	3.1	4.9

Figure 4.4 Average of Averaging Schemes (COUNTRY)

max UB Highest (ending)	250	400	500	333	500	1005
max UB Highest (starting)	100	55	250	55	50	995
max UB Lowest	30	50	60	17	50	30
min UB	1	1	30	1	1	15
LB	1	1	1	1	1	1
<hr/>						
Digit 1	29.6	29.9	29.3	31.5	30.6	30.6
Digit 2	18.4	17.7	16.4	17.8	17.6	19.2
Digit 3	13.4	13.1	12.6	12.6	12.7	13.3
Digit 4	10.2	10.3	10.1	9.7	9.9	9.8
Digit 5	8.0	8.1	8.4	7.8	7.9	7.5
Digit 6	6.5	6.6	7.1	6.4	6.6	6.1
Digit 7	5.4	5.5	6.1	5.4	5.6	5.1
Digit 8	4.6	4.7	5.3	4.7	4.8	4.4
Digit 9	3.9	4.1	4.7	4.1	4.2	4.0

Figure 4.5 Average of Averages of Averaging Schemes (GLOBAL)

a global digit distribution of address data (say for U.N. statistical bureau), and so naturally, the next higher-order scheme is the averaging the averages of the averages; a few such computer calculations are shown in Fig. 4.5.

Results here are even closer to the logarithmic, even though we have strayed far away from the safety of unbiased IPOT borders. Again, one can claim that global data represents something ‘more random’ in a sense than data for just one single country.

The same rationale that propelled us to progress from city to country and then globally necessitates the continuation to higher and higher-order averaging

schemes, especially if we wish to incorporate other types of more complex data as opposed to merely data of street addresses. As expected, the higher the order of the scheme the closer results are to the logarithmic and convergence is rapidly achieved. Interestingly, as order increases, results become completely independent of the values for the limits chosen, be they IPOT or non-IPOT numbers. This is a crucial feature lending more credibility and robustness to the whole averaging scheme theory developed here. Whatever limits are chosen, the logarithmic is found. In other words, the scheme is totally independent of parameters in the limit as higher and higher averaging orders are employed.

In Kossovsky (2006) the author proposes such an averaging approach to Benford's Law. Flehinger (1963) presented rigorous mathematical proof that an exactly such iterated averaging scheme for the integers on the number line approaches the logarithmic in the limit as the number of such iterations goes to infinity. The above discussion could serve as the conceptual framework for going in such roundabout way of calculation and having the logarithmic then emerged; and that such an abstract iterative scheme could refer also to a concrete usage of numbers in the real world, and not merely to the number line itself.

A certain dichotomy, though, exists between the model of simple averaging scheme and actual city address data, and which was overlooked earlier. By taking the simple unweighted average of digit distributions of all the streets within a city, we allocate each street equal importance and weight. Yet, this does not seem like an equitable approach if some streets are long and some are short. There are many more houses on long streets than on short ones, and thus the postman surely must be using house numbers pertaining to long streets much more often than those pertaining to short ones. Taking this fact into account would result in a milder digit distribution that is by far less skewed in favor of digit 1. For example, if a Greek-Parable-like scheme represents street address in a city for which all streets start at 1; the shortest being just one house long, with the house standing alone; the longest street having nine houses; yielding 45 houses in total on all nine streets ( $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 = 45$ ) then direct calculations of digit distribution for those 45 houses still yields a decisive advantage for low digits, but distribution shows a linear-like fall in digit distribution of  $\{20.0, 17.8, 15.5, 13.3, 11.1, 8.9, 6.7, 4.4, 2.2\}$ . Yet, the simple averaging model is still appropriate in the case where a selected group of people (or simply houses) are chosen with the explicit assumption, for one reason or another, that they all live on different streets, in which case equal weight for all streets is a valid assumption. A random pick of, say, 200 people according to last name alphabetical list, age, or income level from a

mega city like New York is almost certain to correspond to 200 different streets. A straightforward way to avoid this dichotomy is to assume that short streets occur with a higher frequency than that of long streets, and that this exactly compensates for their lack of houses, an assumption which seems to be a reasonable one. Certainly there are many more streets with, say, 30 or 200 houses than streets with 35,000+ houses like 5th Avenue or Broadway in New York City. Yet what is difficult to argue here in terms of real-life occurrences is that streets with a single house in them are the most numerous and come with the highest probability. For example, to eliminate this dichotomy for a Greek-Parable-like scheme representing street addresses, it would be necessary to assume that there are nine times as many streets with one house than those with nine houses. That is:

Probability(street with 1 house) =  $9/1 * \text{Probability}(\text{street with 9 houses})$ ,  
and that:

Probability(street with 2 houses) =  $9/2 * \text{Probability}(\text{street with 9 houses})$ ,  
and that:

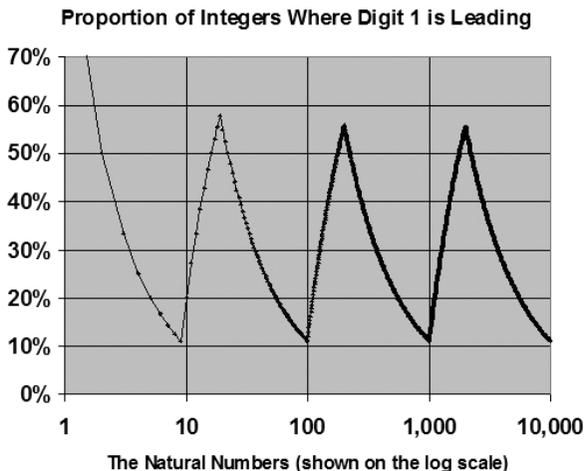
Probability(street with 3 houses) =  $9/3 * \text{Probability}(\text{street with 9 houses})$ ,  
and so forth.

The essential aspect of the simple averaging scheme is incorporating multiple intervals that are growing in focus and having different lengths (UB), yet with all fastened to the same much lower LB, typically 1. Calculations of schemes that are designed totally differently, and where intervals come with fixed length (so that  $UB_i - LD_i = \text{constant}$ , for all of them) but which differ in their position (i.e. shifting each interval one integer forward to a higher location on the  $x$ -axis) all lead to digit distribution that does not resemble the logarithmic in the least, no matter how we vary (in unison) those upper and lower bounds of fixed distances. Fundamental to Benford's Law is the fact that usage of numbers has one unique direction: up! Any data set regarding a particular topic containing in part numbers on, say, (600, 700) where digit 6 leads, would most likely also contain numbers on, say, (500, 600) as well where digit 5 leads (assuming lower bound is somewhere near the origin); whereas the converse is not true. This example shows how the high digit 6 frequently must share its spoils with its lucky lower digit neighbor 5, while the converse is not true. This argument can be generalized to any two competing adjacent digits, hence explaining in general why we do get monotonically decreasing digit distribution. In other words: **To get to high digits one must 'pass' through low digits, and this in essence is what deprives high digits of equality in leadership.**

## DIGITAL PROPORTIONS WITHIN THE NUMBER SYSTEM ITSELF

Let us temporarily put aside the issue of how numbers are used and occur in the real world, and instead focus on digital proportions within the number system itself in the abstract, namely integers on the number line. Figure 4.6 illustrates how proportion of digit-1-led integers within the natural numbers (beyond 10) continuously oscillates from a high of around 55% to a low of around 11%. A log scale is used to facilitate visualization. For example, if the range of integers is from 1 to 3, the first digit 1 occurs one third of the time, or 33.3%. If the range is 1 to 9, the first digit 1 occurs one ninth of the time, or 11.1%. If the range is 1 to 19, the first digit 1 occurs 11/19 of the time, or 57.9%, and so on.

The infinite subset of the even numbers  $\{2, 4, 6, 8, 10, \dots\}$  has density  $1/2$  considering 'all' natural numbers. The infinite subset of the prime numbers  $\{2, 3, 5, 7, 11, 13, \dots\}$  has zero density considering 'all' natural numbers, but has



**Figure 4.6** Portion of Digit 1 Leading from 1 up to a Natural Number N

non-zero density considering a finite set. By the term 'density' here we mean the concentration or proportion of the given sub-set within the entire range of the natural numbers. Yet, the subset of all natural numbers with leading digit  $d$  has no definite density within the set of 'all' the natural numbers. If we stop arbitrarily anywhere on the number line and calculate (from 1) the ratio of integers with leading digit  $d$  (to the left of the 'stop'), the probability value oscillates without convergence. The density of numbers with digit 1 leading oscillates repeatedly between two values which gradually converge to the lower and upper limits of  $1/9$  and  $5/9$  (11.11% and 55.55%). Other digits oscillate between different probabilities in a similar fashion. Putting the horizontal axis on a log scale footing has its price, as it masks and obscures the fact that there is by far more concentration of values towards the right side than the left side progressively for all intervals bordered between integral powers of ten.

The points where digit 1 achieves its maximum potential are: 1, 19, 199, 1999, etc., corresponding to the proportions of  $1/1$ ,  $11/19$ ,  $111/199$ ,  $1111/1999$ , and so forth. The points where digit 1 is at its lowest potential are: 9, 99, 999, 9999 etc., corresponding to the proportions of  $1/9$ ,  $11/99$ ,  $111/999$ ,  $1111/9999$ , and so forth. The simple average between the two consecutive such extremes of  $1/9$  and  $5/9$  yields  $6/9$  or 0.333, not the logarithmic!

The points where digit 2 achieves its maximum potential are: 2, 29, 299, 2999, etc., corresponding to the proportions of  $1/2$ ,  $11/29$ ,  $111/299$ ,  $1111/2999$ , and so forth. The points where digit 2 is at its lowest potential are: 1, 19, 199, 1999, 19999, etc., corresponding to the proportions of  $0/1$ ,  $1/19$ ,  $11/199$ ,  $111/1999$ , and so forth. The simple average between two consecutive such extremes yields 0.213 in the limit.

The points where digit 9 achieves its maximum potential are: 9, 99, 999, 9999, etc., corresponding to the proportions of  $1/9$ ,  $11/99$ ,  $111/999$ ,  $1111/9999$ , and so forth. The points where digit 9 is at its lowest potential are: 8, 89, 899, 8999, etc., corresponding to the proportions of  $0/8$ ,  $1/89$ ,  $11/899$ ,  $111/8999$ , and so forth. The simple average between two consecutive such extremes yields 0.062 in the limit.

The above chart has nothing to do with typical everyday data; it only pertains to our abstract number system itself. One may be tempted to argue that finding here some appropriate algorithm to calculate or arrive at some average would amount to a proof of sorts or a valid demonstration that digit 1 comes with a certain probability, but any such result would pertain only to pure numbers, not

to usage or occurrence thereof. Worse yet, if an algorithm here resulting in  $\text{LOG}(1 + 1/d)$  say is believed, then Benford's Law must be a universal rule, as it is derived from the number line itself pertaining to everything, yet we know that there are plenty of data types and processes that do not obey the law! Some integration techniques here do lead to the logarithmic, others do not. Our complex-iterative averaging scheme above does lead to the logarithmic, and it is accepted as a potential explanation for particular data types only since it can offer real-life, flesh-and blood context such as address data. Stigler's Law, or the simple averaging scheme for a single city pertains to a real city with street configuration having UB varying between IPOT values, and as such it is accepted as a narrow explanation within that context. Flehinger's algorithm in the abstract could not amount to a real-life explanation of the law. Moreover, her iterative scheme surely cannot certify Benford's Law as being universal, pertaining to all data types, even though one might passionately argue that her proof has its strong basis in none other than the number system itself! The author wishes to convey his strong sense of affinity and rapport with the late Betty Flehinger for thinking along almost the same lines in the quest for an explanation of Benford's Law.

## CHAINS OF DISTRIBUTIONS

---

---

In Kossovsky (2006) the author suggests an alternative point of view of the averaging schemes (and, ultimately, that of Flehinger's iteration as well), putting them in a continuous statistical framework. The concept of a chain of distributions is introduced, where the parameter of one statistical distribution is randomly drawn from another statistical distribution. Traditionally parameters of distributions have been thought of exclusively as constants, fixed by the particular nature of the data or random process on hand. Yet, since almost all chains of distributions are indeed logarithmic, this alternative approach to distributions ultimately leads to a better understanding of Benford's Law and so it shall be pursued. The chain idea initiates the study of the digital behavior of such complex statistical constructs of interdependencies, constructs which might find applications in other contexts and disciplines outside the field of Benford's Law and digits.

The alternative point of view of the **simple** averaging scheme involves a two-pronged interpretation of the process. First, each interval itself (relating to a particular  $UB_i$ ) is to be viewed as a continuous random variable, and not merely as a set of integers, utilizing the Uniform distribution on  $(0, UB_i)$ , since each integer was accorded equal probability within its interval. Second, those varying upper bounds ( $UB_i$ ) are viewed as originating from a random process as in the Uniform distribution on  $(\min UB, \max UB)$ , since each interval was accorded equal importance within the whole scheme. Therefore  $UB_i$  are randomly drawn from  $Uniform(\min UB, \max UB)$ . Put another way, instead of attempting to record digital distributions of a variety of intervals with distinct upper bounds, culminating in the averaging of all of them, the corresponding view here is to take directly the digital pulse of the continuous random variable  $Uniform1(0, K)$ , where  $K$  itself is a random number drawn from another continuous Uniform distribution  $Uniform2(\min UB, \max UB)$ . We utilize 0 as LB instead of 1 since this is most typical in the applications of the Uniform. It is noted that  $Uniform2$  exists solely in order to provide  $Uniform1$  with a random parameter  $K$ . To simplify matters,  $\min UB$  is chosen to be 0. Schematically this entire construct is then written as

**Uniform(0, Uniform(0, C))**. Here, the continuous uniform is used, as opposed to the discrete uniform as in the averaging schemes where only integers were considered. Also conveniently, the ranges for the chains start from 0 as opposed to the usage of 1 which was the LB in all the averaging schemes, yet this difference is disregarded and overlooked as it is quite inconsequential when C is large enough and having a sufficiently large order of magnitude. Thus it is argued that these types of **chains of distributions are essentially mirror images of the averaging schemes**. Note that there is actually a fundamental difference between the chains  $U(0, U(0, C))$  and  $U(1, U(1, C))$  whenever C is not large enough, and this is so because (0, 1) is a very special range in the context of Benford's Law as it contains infinitely many IPOT numbers as well as an infinite log spread (thus order of magnitude is infinitely large on (0, 1), strongly supporting logarithmic behavior there).

Computer simulations yield distinct digital distributions, depending on the particular value of C, as might be expected. Of note here is that a choice of 20, say, for C yields the same result as a choice of, say, 200,000. Also of note here is that the value of 100 for C, as well as all other IPOT numbers, yield results that are very close to Stigler's law. The table in Fig. 4.7 shows nine such simulation results.

The next natural step is to pick up numbers randomly from the Uniform on (0, B), where B is being picked randomly from the Uniform on (0, C), and where C is randomly chosen from the Uniform on (0, R). Schematically this is written as **U(0, U(0, U(0, R)))**. This three-sequence simulation should mirror the average of averages scheme (country), and indeed digital results are very much compatible

Value of C:	100	200	300	400	500	600	700	800	900
Digit 1	24.2	31.7	36.7	34.9	33.6	31.0	28.6	27.7	25.7
Digit 2	18.2	12.1	16.1	20.4	20.6	21.0	20.8	20.2	19.1
Digit 3	14.5	11.1	9.1	10.8	13.6	14.6	15.6	15.1	15.3
Digit 4	11.9	9.6	7.6	7.2	7.7	10.4	10.6	11.2	11.6
Digit 5	9.5	8.9	6.8	5.9	5.8	6.3	7.7	8.6	8.8
Digit 6	7.7	7.8	6.7	5.8	4.9	4.5	5.2	6.5	7.5
Digit 7	6.0	7.0	6.3	5.4	4.6	4.5	4.0	4.1	5.3
Digit 8	4.6	6.3	5.6	4.7	4.7	3.9	4.0	3.4	3.6
Digit 9	3.4	5.5	5.2	4.9	4.6	3.8	3.5	3.1	3.1

Figure 4.7 Digital Results from the Chain Uniform(0, Uniform(0, C))

with each other. This three-sequence chain yields better conformity with the logarithmic than for the two-sequence chain. Also, the choice of the last constant R is less crucial to the resultant digital distribution than the choice of C in the previous simulations of the two-sequence chain, where changes in C set off larger swings in digital distribution. But clearly the value of K is still playing an important role here.

Results for the next logical step, namely the four-sequence chain **Uniform(0, Uniform(0, Uniform(0, Uniform(0, M))))**, standing for the average of averages of averages scheme (global), came out very close to the logarithmic as can be seen in Fig. 4.8. Now the value of the constant M matters much less. The extremely close result (shown in the column next to BEN) of the average of these nine simulations to the logarithmic is quite striking! Since these nine simulations are all about different M values, one would expect to get almost perfect agreement with the logarithmic if just one more sequence of parametrical dependency is added, as in  $M = \text{Uniform}(0, 999)$  for example.

Clearly, what is needed here is an infinite-sequence chain to obtain the logarithmic, a requirement that mirrors Flehinger's call to infinitely iterate those averages. It is noted that with each higher-order chain the actual value of the last constant becomes less important, and digit distributions become more uniform (and logarithmic) regardless of the particular choice of that constant.

The conceptual point of view that might be taken here in a general sense is that this result shows that the process of picking a number randomly from a truly random interval, one with an infinite successive arrangement of random upper

M:	100	200	300	400	500	600	700	800	900	AVG	BEN
Digit 1	31.3	29.9	29.5	29.9	29.6	29.8	30.9	31.0	31.1	30.3	30.1
Digit 2	17.5	18.1	17.5	16.9	17.7	17.1	17.7	17.8	18.0	17.6	17.6
Digit 3	12.0	13.0	13.2	12.3	11.7	12.6	11.8	12.3	12.4	12.4	12.5
Digit 4	9.5	9.9	9.7	10.2	9.6	9.9	9.4	9.1	9.0	9.6	9.7
Digit 5	7.8	7.9	8.0	8.1	7.8	8.2	7.2	7.5	7.6	7.8	7.9
Digit 6	6.2	6.3	6.8	6.9	6.8	6.8	6.8	6.7	6.7	6.7	6.7
Digit 7	5.9	5.8	5.7	6.0	6.5	5.9	5.9	5.7	5.9	5.9	5.8
Digit 8	5.1	4.8	5.0	5.4	5.4	5.2	5.6	5.4	5.1	5.2	5.1
Digit 9	4.7	4.3	4.6	4.3	4.9	4.5	4.7	4.6	4.4	4.5	4.6

Figure 4.8 Digits of the Chain Uniform(0, Uniform(0, Uniform(0, Uniform(0, M))))

bounds yields exactly the logarithmic. The ‘very random nature’ of how numbers are being picked here, not even specifying the range clearly, but rather by continuously referring upper bound to yet another random process whose range is again being referred to another one and so forth, suggests the vague and mathematically undefined concept one may call ‘**super random number**’, and which seems to be closely associated with logarithmic behavior.

The Uniform( $a$ ,  $b$ ) has two parameters:  $a$ , the lower bound which we insist on fixing at 0, and  $b$ , the upper bound which we keep referring to another Uniform distribution of the same type. What about other distribution forms? Is the logarithmic convergence of an infinite chain all the about the distribution being Uniform? The answer is decisively in the negative, and the conjecture here is that this digital convergence in tying up parameters in a chain of dependencies is not at all unique to the Uniform distribution, but rather that it is by far a very general principle. Instead of employing just the Uniform, one could select numbers randomly from (almost) any distribution, and with **all** its parameters chosen randomly from other distributions (of the same/different type), whose parameters in turn are also derived from yet other distributions, and so forth. This algorithm should yield the logarithmic distribution so long as there are enough sequences in the chain (infinitely many sequences in principle). Furthermore, as a result of numerous empirical experimentations with several such computer simulations, it appears that convergence is relatively rapid and that after three or four sequences of distributions we come very close to the logarithmic. In practical terms, the fact that chains quickly converge is quite significant, and the expectation is that all physical processes that fit the chain model are nearly logarithmic, even with only two or three such sequences of dependencies, and certainly with four, five, or more sequences. This more general process then should be the one earning the term ‘super-random number’, while Flehinger’s Uniform-like scheme is just one very special case of this much more general principle.

As a notable demonstration of the very general nature of the **infinite chain conjecture**, simulation of a four-sequence chain of distributions is performed as follows: the Normal distribution is simulated where: (A) the mean itself is simulated from the Uniform with parameter  $a = 0$  and parameter  $b = \text{chi-sqr}$ , whose degrees of freedom (d.o.f.) is derived by chance via the throw of a dice; (B) the standard deviation (s.d.) of the Normal is also not fixed but rather simulated from the Uniform(0, 2). It is noted that here four parameters are being chained, while three others are stuck as a constant. Schematically this is written as: **Normal(Uniform(0, chi-sqr( $a$  dice)), Uniform(0, 2))**. Computer

simulations of this short four-sequence chain yield first-digit distribution of  $\{29.9, 18.0, 13.7, 10.5, 8.0, 6.2, 5.5, 4.4, 4.0\}$ , and so agreement with the logarithmic distribution is quite strong, even though there are only four sequences in the chain! [A note on notations regarding the Normal Distribution in use throughout the book:  $\text{Normal}(p, q)$  signifies the Normal with mean  $p$  and standard deviation  $q$ , namely  $\text{Normal}(\text{mean}, \text{standard deviation})$ , as opposed to a different convention in the literature signifying the variance as the second parameter.]

It is tempting and quite intuitive to require all parameters to be chained, leaving none stuck as a constant. In a sense, we may think of the primary distribution on top (the first one, whose existence does not serve as a parameter for another) as a distribution with some deep roots of uncertainty, one that is most thoroughly random, having such uncertain and shifting parameters. For distributions with more than one parameter, the general image of the infinite chain is of a pyramid-like arrangement of densities, with the primary distribution sitting on top, an ever widening base of distributions below supplying the parameters, the parameters of the parameters, and so forth. Figure 4.9 depicts one such possible chain of four sequences. The last row at the bottom contains traditional distribution with fixed parameters, terminating all dependent relationships.

Yet, this simplistic stipulation turned out to be wrong as it depends on the exact nature of parameters/distributions in question. In some cases this stipulation is true, yet in others this is not the case, as some parameters are totally or partially indifferent to chaining and do not contribute to any significant logarithmic convergence. All typical/standard distributions, though, do converge to the logarithmic given that all parameters are blindly and mindlessly (regardless of types) fastened to others in chains, except in a very few extremely rare cases of unusual and odd distribution forms that are totally immune to chaining. For a two-parameter

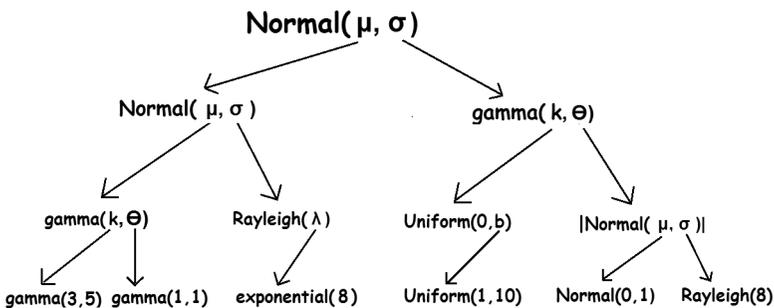


Figure 4.9 Pyramid-Like Arrangement of Distributions Forming a Chain

distribution for example, there are three scenarios: (A) both parameters need to be chained in order for convergence, (B) only one needs to be chained while the other one is totally irrelevant, (C) nothing works here and even if both are chained we do not see any convergence whatsoever (extremely rare). Hence, the chain conjecture is not a universal rule but rather just a very general one restricted to particular types of distributions/parameters. Careful and tedious investigations of digital chain behavior revealed roughly the following principle at work here: scale and location parameters are chainable, while shape parameters are not. In reality the full account of the digital behavior of chains of distributions is a bit more complex. For the sake of better flow and exposition of the whole story of Benford's Law, detailed explanations and results regarding chains of distributions will be given in two separate chapters in Section 6.

In Kossovsky (2006) a second conjecture regarding the chains of distributions is made pertaining to the shortest possible chain having only two sequences, as opposed to the chain with infinitely many sequences — the other extreme in length. The conjecture claims that any two-sequence chain in which all parameters of the primary density are derived from logarithmic distributions is logarithmic there and then without any need to expand infinitely. In other words, if the distributions serving as parameters are logarithmic in their own right, then the chain as a whole is also logarithmic, regardless of length. In symbols:  $\text{AnyDensity}(\text{AnyBenford}) = \text{Benford}$ . Numerous simulations of such short chains (all) confirmed the second conjecture, which is true exactly for the very same types of parameters and distributions that the first (infinite) conjecture refers to.

An example of the second conjecture is the Uniform distribution with parameter  $\mathbf{a}$  fixed at 0, and with parameter  $\mathbf{b}$  being a random pick (with uniform chance) from the exponential growth series  $8 \cdot 3^N$  for values of  $N$  in  $\{0, 1, 2, \dots, 5000\}$ , thought of as a data set. As seen earlier, this exponential growth series is nearly perfectly Benford if taken far enough, hence so is the whole chain.

**In symbols: Uniform(0,  $8 \cdot 3^N$  for  $N = \{0, 1, 2, \dots, 5000\}$ ).**

There are three conceptual points of view regarding chains of distributions. The illustrative example of Uniform1(0, Uniform2(0, Uniform3(0, C))) leads to three different interpretations:

- 1) **Statistical Theory:** Distribution Uniform1 does not have the traditional fixed  $\mathbf{b}$  parameter; rather, it has a variable  $\mathbf{b}$  parameter, namely Uniform2. Likewise, distribution Uniform2 has its variable  $\mathbf{b}$  parameter as Uniform3.

- 2) **Stepwise Process of Value Realization and Parameter Insertion:** An actual number called  $N_3$  is simulated (realized) from  $\text{Uniform3}(0, C)$ . That single number is then inserted into  $\text{Uniform2}$  as the traditional fixed  $b$  parameter.  $\text{Uniform2}(0, N_3)$  is then simulated to yield an actual number called  $N_2$ . The value  $N_2$  is then inserted into  $\text{Uniform1}$  as the traditional fixed  $b$  parameter.  $\text{Uniform1}(0, N_2)$  is then simulated to yield the final value called  $N_1$ .
- 3) **Graphical; Second-Generation Distribution:** Turning the charts of multiple density curves into one large and singular resultant curve in the usual scatter plot of  $\text{PDF}(X)$  versus  $X$ . For simplicity, let us limit the focus to a very short and simple chain having only two sequences. As an example, we consider the short chain  $\text{Uniform1}(0, \text{Uniform2}(0, 10))$ . Such basic/atomic chain could be considered as the aggregation of, say, 10 separate density curves all spread on the interval  $(0, 10)$ , as in  $U(0, 1)$ ,  $U(0, 2)$ ,  $U(0, 3)$ , ...,  $U(0, 10)$  combined together. It requires giving each mini-distribution equal weight, combining all 10 of them as one large distribution curve. If one is inclined to use histograms to represent data, then here we simply create a sample histogram for each of the 10 Uniforms separately via simulations (number realization), utilizing the same number of histogram-values for all of them — say, 10,000 (to give all ten Uniforms equal importance) — and then plot all these numerous values ( $10 \times 10,000$ ) as the final histogram of the whole chain on a fresh and much larger sheet of paper. The same concept applies to probability density curves when these need combining and to second-generation distributions in general. Needless to say, this set of 10 is not sufficient. All those infinitely many in-between Uniform such as  $U(0, 0.005)$ ,  $U(0, 4.8387)$ ,  $U(0, 8.962)$ , and so forth should also be included, although a finite and discrete collection of such distributions in steps of, say, 0.01, should be more than sufficient in practical terms. The aggregation of all these individual densities is then simply a singular density curve in the classic sense, representing a unique distribution in its own right. It must be noted that not only in the simple two-sequence case above, but even in the infinite chain case, the ultimate distribution to consider and to ‘draw’ is the one sitting on top (not supplying parameters), albeit with numerous versions of it, each with a slightly different parametrical value, and with all versions having the same distribution form. For example, the three-sequence chain  $\text{Normal}(\text{Uniform}(0, \text{exponential}(8)), \text{gamma}(7, 5))$ , is ultimately nothing but many versions of the Normal distribution superimposed.

Does the chain of distribution have anything to do with real-life data? In other words, could it constitute an explanation for the logarithmic behavior of (at least)

some data types? We have already considered global address data, and seen how it can be modeled on averaging schemes, which in turn nicely correspond to chains of Uniform distributions. Surely other data types that are similar to the address data in structure could also be modeled as such and thus come under the protective umbrella of the chains. Could the chains also constitute an explanation for the logarithmic behavior of Aggregate Global Data? Here, the answer is not very certain; it is not at all straightforward to argue that the chains could serve as a model for AGD, except in pointing out the fact that in extreme generality lower bounds of AGD tend to be ‘fixed’ around 0, 1, or some very low number, while upper bounds surely tend to vary wildly upwards towards very high orders of magnitudes, with all this resembling a great deal the spirit of the (simple) averaging schemes. Besides, trade-offs and cancellations of digital forces working in opposing directions can perhaps explain away deviations of AGD from the pure and generic structure of the averaging schemes. In reality, the explanation of AGD falls under the realm of Hill’s super distribution, which will be discussed in the next chapter.

A stronger argument for the relevance of the chains to real-life data could be made if we ponder the preponderance of causality in life, the multiple interconnectedness and dependencies of entities that we normally measure and record, and that so many measures are themselves parameters for other measures. This could be the rationale for how the chain of distributions may explain a good portion of physical data, such as geological, astronomical or chemical data sets. A detailed conjecture along these lines of thought will be made in a later chapter in Section 5 applying chains of distributions as the explanation for the single-issue physical manifestations of Benford’s Law.

Steven Miller (2008) gave a rigorous mathematical proof of the first- and second-chain conjectures for the particular one-parameter distribution cases where PDF is defined exclusively on  $(0, +\infty)$ , such as  $\text{Uniform}(0, b)$ ,  $\text{exponential}(\rho)$ , and  $|\text{Normal}(0, \sigma)|$ .

## HILL'S SUPER DISTRIBUTION

---

---

More than half a century after the re-discovery of the phenomena in 1938 by Frank Benford and after numerous failed attempts at proofs, plagued by suspicion and enduring terms such as numerological, mystical, and pseudo-mathematical, finally a rigorous mathematical explanation for one very important class of data was given in 1995 by Theodore P. Hill, demonstrating that a second-generation distribution consisting of infinitely many distributions — all defined on the positive  $x$ -axis — is logarithmic in the limit as the number of distributions approaches infinity. The distributions being aggregated are allowed to take on almost any form, be it the Uniform, the Exponential, the Gamma, and so on. They are also allowed to assume any value(s) for the parameter(s), as long as defined range is strictly positive. As a number realization scheme for Hill's super random number, the first stage is to randomly select the form of the distribution, and then to choose the value(s) for the parameter(s). The second stage is to randomly simulate an actual number from the chosen distribution. Those two stages are repeated over and over again to generate more values, and it is logarithmic in the limit as the number of realizations goes to infinity. In practical terms convergence to the logarithmic is encountered much sooner, say after 500 or 1000 realized values from distinct distributions, depending on the desired level of accuracy or conformity. With the exception of data sets containing negative values, Hill's result is tailor-made for the Aggregate Global Data Interpretation of Benford's Law, since AGD is nothing but a composition of a huge variety of data sets corresponding to random variables and distributions. Since the vast majority of values in typical everyday use are positive, his model is almost perfectly fitting. The remarkable feature of this result is that the distributions do not have to be (actively) fixed or anchored near 0 or 1, nor do they have to gradually increase their focus by varying their upper bounds in any way (as was necessary for the averaging schemes). Yet it must be acknowledged that common distributions such as Normal(0, 1) and Uniform(-10,+10) are not allowed in the model, since they are defined in part over the negative side of  $x$ -axis.

Surely Hill's random number from his distribution of all distributions can be viewed as some kind of a super random number, a number with deep roots of uncertainty, just as was said about the infinite chain of distributions. Yet, there are two important differences. The first is that in Hill's scheme the distributions are independent of each other, while in the chain scheme the distributions are considered sequentially, having a certain structure and order as they depend on one another. The second difference is that for the chain scheme, the final outcome is a collection of distributions of the same form (that of the top distribution not serving as a parameter for any other distribution), differing only with respect to their parameters, while in Hill's scheme it is a diversity of forms. It is worth noting that Hill's super distribution is parameter-independent, or parameter-less, just as the infinite order averaging schemes and the infinite chains of distributions are.

Another crucial difference between Hill's super distribution and the chain of distributions is applicability. Hill's model can account for the combined real-life typical data sets in the form of Aggregate Global Data, while the chain is able to explain other data types fitting its model (such as address data and some single-issue physical data sets for example.)

Putting Hill's super distribution under the microscope, one sees numerous two-sequence chains of distributions hiding there, stubbornly avoiding the lime-light. Moreover, actually that is **all** one could see there — short chains exclusively! To see why, let us take the exponential distribution as one example, one minor item on Hill's overall super distribution infinite agenda. Surely exponential(56.7) is just as likely to be part of his scheme as, say, exponential(0.0139), and so forth. Also, exponentials with parameter on, say,  $[0, 1]$ , are just as likely to be included as exponentials with parameter on  $[1, 2]$ . Therefore, this minor item on his super distribution agenda is simply the two-sequence chain of distribution **exponential(uniform(0,  $+\infty$ ))**! The same argument follows for the Uniform, the Gamma, and all the other relevant distributions. The fact that these two-sequence chains of distributions are quite close to the logarithmic in their own right, may constitute partial explanation of the mechanizations at work here. The contents of Chapters 102, 103, and 107 relate to this possibility.

Hill's model indirectly leads to an advantage for low digits and to the logarithmic distribution simply due to his restriction on all distributions to be defined exclusively on the positive  $x$ -axis. Even though there is no direct or active anchoring of distributions' lower bound at 0 or at 1, as they are free to start anywhere on the positive  $x$ -axis depending on form and parameter, yet, the totality of all the

distributions as a single aggregated density tends to crowd out naturally near the origin/zero, since this is a definite and obvious limit for all lower bounds! Upper bounds on the other hand are unbounded in the positive infinite direction; hence we get again that dichotomy between LB and UB and the usual stretching of UB seen earlier in the averaging schemes. It is this dichotomy between LB and UB that appears in the chain of distributions as well as in Hill's super distribution that constitutes the common denominator between them, resulting in logarithmic behavior for both.

In other words, by creating this '**wall**' at the origin, Hill indirectly piles up relatively larger portion of that aggregate density near the origin, and lets the resultant tail to the right fall continuously and steadily as focus moves in the positive  $x$ -axis direction. Such a vista of Hill's model suggests a radical generalization of it, namely the distribution of all distributions defined over  $(+IPOT, +\infty)$ , such as, say,  $(+10, +\infty)$ . In other words, since 0 can be viewed as just another IPOT value, namely  $10^{-INFINITY}$ , the restriction of values being strictly positive can be interchanged for being strictly over a particular IPOT value. Future mathematical research may hopefully one day confirm this outlandish conjecture.

Hill's model cannot be a valid explanation though for the (almost) perfectly logarithmic data sets regarding single-issue physical phenomena, such as earthquake depth and time between occurrences, river flow, population count, pulsar rotation frequency, half-life of radioactive material, and so on. His model does not show any immediate or obvious relation to the logarithmic way Mother Nature generates her physical quantities. Single-issue natural phenomena represent something quite striking in Benford's Law and require a radically different explanation than that supplied by Hill's model, where data was found to be logarithmic following our man-made and artificial aggregation, compilation, and mixing. It is very hard or rather impossible to argue that river flow or earthquake depth are the results of some aggregation of numerous 'other' invisible and mysterious mini distributions.

## THE SCALE INVARIANCE PRINCIPLE

---

---

A significant step forward, and a renewed mathematical interest in the leading digits phenomenon in general, came about with Roger Pinkham's 1961 seminal paper attempting to mathematically and rigorously prove Benford's Law. Additional mathematical pioneering work by Ralph Raimi in the late 1960s further propelled the field into mathematical and scientific respectability. Pinkham's novel approach in leading digits is based on the scale invariance principle; other such attempts are based on the base invariance principle. Scale and base invariance arguments claim that if there is indeed any universal law for significant digits, then it should be independent of units and scales as well as the base in the number system employed by society, because scales and base are cultural, arbitrary, and do not represent any fundamental properties of numbers or nature. In other words, had Benford's Law been dependent on humanity using the French meter as opposed to the British feet, the law could not be considered universal. In the same vein, had Benford's Law been dependent on humanity using Base 10 number system as opposed to, say, Base 8 system, then the law could not be considered a universal rule. The general first digit Benford's Law is in indeed base-invariant, since it is written as a single algebraic expression for all bases:  $P[d \text{ is first}] = \text{LOG}_{\text{BASE}}(1 + 1/d)$ . For example, for societies using Base 6 number system with  $\{0, 1, 2, 3, 4, 5\}$  as the set of all possible digits, Benford's Law states that  $\{38.7\%, 22.6\%, 16.1\%, 12.5\%, 10.2\%\}$  are the probable proportions in logarithmic data sets for the first digits. This vector of proportions is calculated from the general expression above as  $\{\text{LOG}_6(1 + 1/1), \text{LOG}_6(1 + 1/2), \text{LOG}_6(1 + 1/3), \text{LOG}_6(1 + 1/4), \text{LOG}_6(1 + 1/5)\}$ .

Ultimately, Pinkham's explanation of the phenomena was not accepted, although it generated a considerable appreciation of this very unique mathematical aspect of the law, being the only distribution that is scale-invariant. This fact guarantees that no potential competitor to the logarithmic (another distribution) will ever emerge any time in the future. Pinkham's scale invariance principle (plus base invariance) plays an essential role in Hill's mathematical constructions in his proof. In any case, the difficulty with Pinkham's argument is that Benford's Law was

indeed about everyday and scientific data as we have it in our society, not about our number system or pure numbers; that the law could in principle depend on scale; and that his argument relies on the unproven assumption that some scale-invariant first-digit law exists in the first place.

Yet, the scale invariance principle, if assumed, implies the logarithmic distribution! Pinkham has demonstrated that the logarithmic distribution is scale-invariant, and that it's the **only** distribution with such a property; therefore any digit law that is independent of choices of scale must be the logarithmic.

On the face of it, the scale invariance principle appears totally unfounded. For example, people's height measured in feet yields mostly 5 and 6 as the leading digits, while on the meter scale digit 1 takes a very strong lead. Yet, the principle is quite intuitive and even compelling when considered in the context of Aggregate Global Data Interpretation. The rationale for the principle rests on the notion that a change in scale would have such varying and independent effects on all those numerous real-life data sets in terms of leading digits, that it will all add up to nothing. In other words, even though a change in scale revolutionizes digital leadership for almost all data sets, those digital revolutions generally take totally different turns, canceling each other's effects and leaving the net results on overall digital leadership unaffected. For a more specific example, consider two data types: heights of people measured in feet where most quotes are assumed to be on (4.5 ft, 6.7 ft), and heights of buildings in a certain region or country where most quotes are on (65 ft, 193 ft). A change of scale to meters using the conversion formula (feet measurements)\*0.30919 = (meter measurements) would give roughly the new intervals (1.39 m, 2.07 m) and (20.1 m, 59.7 m). Clearly digit 1 gained considerable leadership with people's height due to the scale change, but lost badly with the buildings. This trade-off is due to data representing something real and physical, and that the two intervals in the example above relate to each other in a fixed manner independent of any scale chosen. The moon (radius 1738 km) is smaller than the Earth (radius 6378 km) no matter how we measure length, in kilometers or in miles. Moreover, the relative sizes, namely the ratio 1738/6378, is exactly 0.2725 in any scale system whatsoever. The moon is 27.25% the size of the Earth, and this fact is scale-invariant!

The scale invariance principle implies that transforming any large enough logarithmic data set by multiplying each value by (any) single factor would barely nudge their digital distributions from their near logarithmic configuration. Readers are encouraged to attempt performing this rather striking demonstration

of the scale invariance principle. Since this is a limiting result that is strictly true only for perfectly logarithmic and infinitely large data sets, expectations for such digital stability should be tempered, and in reality some slight deviation in digital configuration always accompanies scale changes. [There exists no computer having a file containing perfectly logarithmic and infinitely large data set!] Also one should always bear in mind that a scale change **does** dramatically affect leading digits distribution of non-logarithmic data sets.

It must be acknowledged that certain data types (such as count data) have nothing to do with scales. Examples of such count data are number of complete cycles, number of accidents per year, number of people in an area (population census), and so forth, with a particular non-optional unit of measurement. The scale invariance principle is not relevant to count data in a theoretical sense, yet it is applicable there as well in the sense that digital configuration still does not change at all almost under any multiplicative transformation of such logarithmic count data. [Instead of calling the conversion of the data “re-scaling”, it is called “multiplicative transformation” — merely a matter of semantics].

The scale invariance principle implies that if data  $X$  is logarithmic then so is the rescaled data  $sX$ , for any value of  $s$ . The logarithmic distribution is the only one having this property. The so-called ‘converse’ of this principle is also true, namely that if data  $Y$  is not logarithmic, then no rescaling could ever make it be so —  $sY$  is still non-logarithmic regardless what value  $s$  takes. In summary, rescaling proved conservative for both logarithmic  $X$  and non-logarithmic  $Y$ . Also of note here is that the principle only says that rescaling is ‘harmless’ to logarithmic behavior given that data is logarithmic to begin with. Yet, the principle does not imply that rescaling is ‘harmless’ to non-logarithmic data as well, in the sense that resultant digital distribution of rescaled data is just as far from the logarithmic as it was prior to rescaling (say if measured via SSD). On the contrary, non-logarithmic data with mild deviation could end up much farther from the logarithmic under a particular rescaling scheme (or it might get much closer). Interestingly, a process of gradual and repeated rescaling of non-logarithmic  $Y$  data set, transforming  $Y$  into  $sY$  for  $s$  in the interval, say, (1, 500) being incrementally increased by 0.01 at a time for example, yields repeated digital cycles above and below the logarithmic.

## PHILOSOPHICAL AND CONCEPTUAL OBSERVATIONS

---

---

At times the reason a particular data set obeys the law comes about as a result of simultaneous influences acting upon it, leading to a convergence. The data might involve some multiplications that are limited and not sufficient for full convergence. It might also be actually a limited mixture of several different phenomena aggregated together, and as such it gets some extra push via Hill's model, but not sufficiently so for convergence as a sole factor. It might also include some aspects of simple or more complex averaging schemes, but possibly lower bounds are not nicely fixed near 0 or 1, or it may be only two sequences deep. The **confluence** of all these factors could then lead nearly to convergence. Such hybrid manifestation of the various explanations of Benford's Law within a single data set must be considered as a distinct possibility, and perhaps it is much more common in real-life data sets than is currently expected or believed. An explicit numerical example of such hybrid causes leading to logarithmic behavior is given in Chapter 95.

The simple averaging scheme relating to Stigler's Law shouldn't be considered a failure, as it actually applies to those statistical processes and variables where LBs congregate around 0 or 1 while UBs vary upwards, and so forth. Each random process or variable has its own unique digital signature, be it the logarithmic signature or another vector of digital proportions. Benford's Law is an example of a digit law, albeit an extraordinarily useful one since it is by far the most common. There are in principle infinitely many different digit laws for the infinitely many other very specific cases of processes and distributions that differ from Benford. As an example of just one such non-logarithmic case, consider the ratio of two continuous random variables that are uniformly distributed on (0, 1), that is:  $\text{Uniform}_1(0, 1)/\text{Uniform}_2(0, 1)$ . Digit distribution of the first order for this process is: {33.3, 14.8, 10.2, 8.3, 7.4, 6.9, 6.6, 6.3, 6.2}, and which can be expressed directly as  $\text{Probability}[\text{1st digit} = d] = (1/18) * (1 + (10/d)/(d + 1))$ . This is but one example of the many specific statistical processes and variables that obey digit laws decidedly different than Benford's.

The Aggregate Global Data Interpretation of Benford's Law relates to the totality of our data, not to any subset thereof. Since the process of gathering numbers from our enormous collection of everyday data is quite time-consuming, any researcher that starts such data gathering methodically by choosing numbers from particular data types (sources) sequentially, spending a lot of time and effort on one type of data first to collect many values, then moving on to the next type, etc. and stops after exhausting a good number of such types because of a time limit, could be severely disappointed by the result. The law requires that this type of data gathering procedure must be carried out to its ultimate end, an impossible task! On the other hand, if the method of picking numbers is such that it constantly shifts the data types (sources) from which it draws numbers, then the logarithmic will be confirmed quite rapidly. In other words, only a single number or very few numbers should be drawn from one source, then quickly moving on to the next source, etc., as opposed to gathering numerous numbers slowly from each source, leaving little time and resources to incorporate a truly large variety of sources. For example, 800 batches of five numbers each where the batches are picked from 800 distinct sources, totaling 4000 values, is strongly in the spirit of Hill's Model, while five batches of 800 numbers each picked from five distinct sources is not so at all. It's the ratio of numbers per source to the number of sources that must be kept low, so that even a large number from each source drawn from a much larger number of sources is acceptable. In any case, if by sheer luck some particular subsets of AGD are intrinsically logarithmic (or very close to it) the first research method may still yield acceptable results.

Benford's Law expresses digital proportions as log values. That the simple proportions of the 'letters of our numerical language' utilize something as 'remote' and 'complex' as logarithms might first appear a bit peculiar or surprising. Yet perhaps it should not be so. Our number system is based on powers as well as additions and multiplications. The number 578 means  $5*10^2 + 7*10^1 + 8*10^0$ . To ask about the value of the log of 578 essentially leads to unifying all these three different powers (0, 1, and 2) and condensing them into one singular power; allowing fractional power; and insisting on the value 1 as the coefficient. Therefore the question of  $\text{LOG}(578)$  leads to  $5*10^2 + 7*10^1 + 8*10^0 = 1*10^L$ , implying that L is 2.762, and therefore we could write  $5*10^2 + 7*10^1 + 8*10^0 = 1*10^{2.762}$ . This reminds us that the idea of logarithm is not very foreign or far detached from the basic concepts of our number system. The only novel or extraordinary part in the concept of logarithm is really the concept of allowing a fractional power.

And the only novelty in a (rational) fractional power is the novelty of the square root, since  $10^{N/D} = D$ th root of  $(10^N)$ , which in turn requires only the simple concept of the inverse of power. The definition of **powers** is the inverse definition of **logarithms**, and vice versa, and both point to the same conceptual construction. The identities  $\text{LOG}(10^Q) = Q$  and  $10^{\text{LOG}(Q)} = Q$  remind us of this intimate connection between them.

In addition, another relevant aspect of our number system worth noting here is that **digital size** of a whole number (integer) is directly related to its **log value**. For example, for a number just below IPOT values, digital size is approximately equal to its log value. Digital size of 97,145 is five, and its log is 4.99, or almost 5. Digital size of 95 is two, and its log is 1.98, or almost 2. For an integer just above IPOT values, digital size is approximately equal to its log value plus one. Digital size of 102,145 is six, and its log is 5.01. Digital size of 105 is three, and its log is 2.02. While digital size connects to the log value of the integer in question, digital proportions in large (logarithmic) data sets are expressed in terms of differences in the log values of all our nine digits.

Which Roman numerals would predominate and which would be relatively rare (had we been still using such an inefficient number system) is something to be investigated. Could a society which utilizes Roman numerals arrive at an exact, stable, and consistent law, therefore enabling itself to detect fraud whenever provided data deviates from the norm? There are seven symbols in the Roman numerals system; {I, V, X, L, C, D, M} signifying the values {1, 5, 10, 50, 100, 500, 1000}. The numbers 1 to 10 are expressed in Roman numerals as in {I, II, III, IV, V, VI, VII, VIII, IX, X}. Surely one would not expect to find there any concise and elegant expressions such as  $\text{LOG}(1+1/\text{symbol})$  when values are substituted with their Romanic symbols. To find a perfect equality of proportions in the occurrences of those Romanic symbols when all real-life quantities are converted into such an antiquated system would not only constitute a remarkable result, but rather a fantastic coincidence bordering on the magical. Nor is there any expectation of symbol equality for the ancient Mayan, Egyptian, and Greek number systems. Our current (decimal) positional number system demonstrates its efficiency and perfection by the simple fact that it hasn't been revised at all in centuries, while our modes of communication, transportation, production, calculation, and others, have all been drastically altered and improved. Evidently, another demonstration of its remarkable efficiency is the existence of Benford's Law, namely its ability to account exactly for the distribution of the proportions of all its symbols (digits) in

such a concise manner as in  $\text{LOG}(1+1/d)!$  Our naïve intuition that all digits should occur with equal proportions is misguided, and is driven by the lack of the realization that any (arbitrarily) invented number system in use for representing quantities out there in the real world should be considered highly successful, novel, and unique if it manages to bring about an equality in the occurrences of its symbols. Unless coincidental by some very rare chance, only the active effort, intensive labor, direct intentions, and ingenuity on the part of that Arabic, Babylonian, or Indian ‘number inventor’ could lead to such an even result of having all numeric symbols occurring exactly with equal proportions! Had that ‘number inventor’ stubbornly insisted on this particular symbol-equality outcome within his or her preferred number system (constructed initially without regards to occurrences of symbols), then it would take some divine intervention to actively ‘adjust’ all the physical quantities out there in the real world in order to achieve such an elegant and even outcome — surely a monumental and staggering task even for the Gods! It is essential to emphasize the fact that digit or symbol distribution is derived from the interaction and combination of two separate factors: (1) the number system in use and its symbols, (2) the way quantities in the physical world are constituted in and of themselves without any reference to a number system. It takes two to tango!

## SOME GENERAL RESULTS

---

---

Phone numbers, lottery numbers, serial code numbers, assigned ID numbers, social security numbers, driver license numbers, passport numbers, index numbers, post office zip code numbers, and such, do not conform to Benford's Law and their digital distributions are for the most part uniform (i.e. with equal proportions). For these types of numbers, each digit is randomly selected with equal probability independently of its adjacent digits, so that digits occur with an equal 10% probability. Put another way, a lottery 'number' with six digits in reality is not a single number as it is commonly perceived, but rather it's a set of six totally independent and unconnected digits, necessitating six different decisions. The common denominator in all of these types of numbers is that they (or any single digit within the 'numbers') do not represent quantities or counts of anything. For example, if Frank's phone number is 2474633 while that of Simon is 9884589, then it would be ludicrous to conclude that Simon (as compared with Frank) is older, heavier, wiser, taller, richer or anything else quantitatively, since these digits do not stand for any quantities in the physical world.

The law rarely applies to data sets where the range is narrowly confined within one or two orders of magnitude. For example, if data lies within (10, 1000) or within (510, 6030), the law is not expected to be observed. On the other hand, for data within (1, 1000000) say the law is usually observed. Consequently, when the numbers are integers and specifically short, being only one- or two-digit long, the law is not expected to be observed, since order of magnitude here between the lowest (1) and the highest (99) is only two.

Artificial numbers that are influenced by human thoughts, such as ATM withdrawal numbers, or a particular price schedule that is made less to reflect cost and more to attract clients, do not obey Benford's Law. Quite often prices are set to fall below a psychological barrier, such as \$4.99 which is perceived by clients as much lower than \$5.00. Typical amounts taken from an ATM machine such as \$20, \$240, \$100, \$60, \$80, and so forth, are too even and particular, invented in the minds of account holders. Another important exception to the law is for numbers with a built-in maximum or minimum value, either artificially human-made or due to some natural barrier to values.

Adhikari and Sarkar (1968) showed that if a variable or a data set  $X$  obeys Benford's Law, then so does the variable or data set  $1/X$  or  $c/X$  for  $c > 0$ . Some have called this property of the logarithmic distribution the '**reciprocal invariance principle**'.

Two important results by Hamming (1970) relate to products and divisions of several data sets or distributions. Let  $X$  be a continuous random variable with an exact logarithmic behavior. Let  $Y$  be any other continuous random variable, logarithmic or non-logarithmic. Then the product  $X*Y$  and the ratios  $X/Y$ ,  $Y/X$  all satisfy Benford's Law as well. One may call these two properties of the logarithmic distribution as the '**multiplicative invariance principle**' and the '**divisional invariance principle**'. The ramification of these two remarkable properties is profound, since it applies to any distribution form  $Y$ ! These two properties may be thought of as 'propagators' of the logarithmic distribution, guaranteeing to spread it around whenever a logarithmic data set gets 'arithmetically connected' with any other data type, with the result that it in turn 'infects' other data sets with the logarithmic configuration, and so on. Moreover, once emerged, it persists! In Hamming's words: "The [logarithmic] distribution persists under multiplication and division and cannot be broken by any choices for the other factors."

An even more remarkable result given by Hamming shows that in long computations "Benford's Law seems to appear out of nowhere!" This statement springs from the fact that products or ratios of independent and continuous random distributions (be they logarithmic or non-logarithmic) converge to the logarithmic in the limit. Let us use formal notation of the above result for clarity: let  $X, Y, Z, Q, \dots$  etc., be any random distributions, logarithmic or non-logarithmic, then either the product  $X*Y*Z*Q* \dots$  etc., or the quotient  $X/Y/Z/Q/ \dots$  etc., is logarithmic in the limit as the number of distributions becomes large. The remarkable feature of this property is that even though none of the distribution is logarithmic (i.e. none is able to infect others), nonetheless the logarithmic property emerges after sufficient numbers of such multiplications or divisions (out of nowhere)!

These properties strongly remind us of the results from the chain of distributions, where the logarithmic was found either (I) when the length of the chain was large, consisting of many sequences (**emerging out of nowhere**), or (II) when the last parametrical distribution at the bottom was logarithmic in its own right (**infecting others**).

As a direct or related consequence of the above results by Hamming, Adhikari, and Sarkar, it is noted that if  $U$  is a random Uniform distribution defined on  $(0, 1)$ , then  $(1/U)^N$ , with  $N$  as an integer, is Benford in the limit as  $N$  gets large. In other

words, repeated multiplications of the reciprocals of  $U(0, 1)$  is a process leading to a logarithmic behavior. Note that only a single distribution of the reciprocal of the Uniform is involved here, not multiplications of diverse distribution types as was originally suggested by Hamming. In any case, once again the logarithmic emerged out of nowhere! Yet, surely this is just one more example of Hamming's result, since it represents constant multiplications of a random distribution, albeit of a singular form and parameter. Adhikari and Sarkar also proved that the same logarithmic result is obtained for repetitive divisions of  $U(0, 1)$ , namely  $U_1/U_2/U_3/U_4/\dots/U_N$ . More generally Adhikari and Sarkar found that for any random variable  $Y$  defined on the positive axis, be it logarithmic or non-logarithmic, and  $U$  being a random Uniform distribution defined on  $(0, 1)$ , repeated division of  $Y$  by successive  $U$ 's, namely  $Y/U_1/U_2/U_3/U_4/\dots/U_N$  is logarithmic in the limit as  $N$  gets large.

Computer simulations strongly confirm that **any** random data set  $X$  (logarithmic or non-logarithmic) transformed by raising each value within the data to the  $N$ th power, converges to the logarithmic as  $N$  gets large. That is, the set  $\{X_1, X_2, X_3, \dots, X_M\}$  of  $M$  realizations from  $X$  transformed to  $\{X_1^N, X_2^N, X_3^N, \dots, X_M^N\}$  is logarithmic as  $N$  gets large, culminating in a near-perfect agreement with the logarithmic in the limit. Here as well, Benford's Law may be said to 'appear out of nowhere'. It should be noted that here intermediate products are discarded, and only the last  $N$ th power is retained and considered, thus this is really not a repeated multiplications process of the exponential growth series type.

It must be emphasized that the above results by Hamming, Adhikari, and Sarkar are all about random processes, and should not be confused with deterministic ones, even though terms such as multiplication, products, and powers are involved. Empirical examinations show that for any variable or data set  $X$  obeying Benford's Law, the transformation  $X^N$ , where each value  $X_i$  is being transformed by raising it to the  $N$ th power, is also logarithmic, provided that  $N > 1$ , namely squares, cubes and higher (integral or fractional) powers. Yet for  $N = 1/2$ , namely square root transformation, often transformed data may still end up quite close to the logarithmic. For  $N = 1/3$ , namely cube root transformation, the transformed data typically (more often than square roots) ends up quite far from the logarithmic. For  $N = 1/4, 1/5$ , and other such higher roots, data does even come close to being logarithmic. In addition,  $\text{LOG}_B(X)$  transformations do not inherit the logarithmic property of logarithmic data  $X$  at all, and this is so regardless of the value of the base  $B$ , be it the base under which digits are defined and examined in the number system in use, or any other base.

The Fibonacci series  $\{1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 233 + 377, \text{etc.}\}$  begins with  $\{1, 1\}$  as two arbitrarily chosen starting points. Subsequent

elements beginning with the third one are simply the addition of the previous two elements, expressed algebraically as  $X_N = X_{N-1} + X_{N-2}$  for  $N = \{3, 4, 5, \text{etc.}\}$ . Interestingly, even though the series is defined only in terms of additions, it approaches approximately a repeated multiplication process very early on, with the golden ratio 1.61803399 as the factor, since successive elements can be approximately obtained by simply multiplying the previous element by the golden ratio. This explains why the Fibonacci series is almost perfectly logarithmic. For example, the eleventh element 89 is approximately equal to the golden ratio times the tenth element 55. That is:  $55 * 1.61803399 = 88.99 \approx 89$ . Algebraically, the golden ratio approximation claims that after the first few elements (say, from the eleventh element on) the series becomes nearly like a standard exponential growth series expressed as  $X_N = 1.618 * X_{N-1}$  or equivalently that after the tenth element of 55, for example, it becomes approximately  $X_N = 55 * (1.618)^{(N-10)}$  for  $N = \{11, 12, 13, 14, \text{etc.}\}$ .

This last expression for all elements after the tenth one can be written more compactly as  $X_N = 0.4473 * (1.618)^N$  for  $N = \{11, 12, 13, 14, \text{etc.}\}$ . The list of all the factors in each step forward within the Fibonacci series, namely the ratios of adjacent elements  $\{1/1, 2/1, 3/2, 5/3, 8/5, 13/8, 21/13, 34/21, 55/34, 89/55, 144/89, 233/144, 377/233, \text{etc.}\}$  approaches the golden ratio quite rapidly after the tenth element or so as seen in the values  $\{1.00, 2.00, 1.500, 1.6667, 1.600, 1.6250, 1.6154, 1.6190, 1.6176, 1.6182, 1.6180, \text{etc.}\}$ . It is also noted that due to its relatively high growth rate of 61.8%, the Fibonacci series passes through an IPOT number each five terms approximately, thus the second requirement of Chapter 20 can be easily waived and we need not worry much where we start and where we end, only that enough elements are considered. As a check:  $1.618^5 \approx 11 > 10$ , so that in about five terms the cumulative increase is over tenfold, and an IPOT number is passed. Surely, all intermediate products of the Fibonacci series (its elements) are kept and considered as to their digital configuration, not merely that 'last' one, and this aspect is crucial for logarithmic behavior as discussed earlier. Finally, it must be noted that the Fibonacci series is a deterministic process; there is nothing random about it.

The dynamical system of the form  $S_{N+1} = (S_N)^2 + 1$ , with the first term  $S_1$  being any arbitrarily selected value is logarithmic. For example,  $S_1 = 3$  yields the infinite series  $\{3, 10, 101, 10202, 104080805, \dots\}$ . Here all intermediate elements are kept and considered, not just the 'last' term. Such a sequence is logarithmic in the limit as  $N$  gets large, and no matter what initial value  $S_1$  takes. This is so in spite of the fact that the series incorporate also some addition, since the

term +1 is negligible relative to the much larger products beyond the starts of a few initial elements of possibly small values. Beyond a certain length, though, the system explodes upwards quite rapidly. It is noted that the series is not considered a random process in any way, but rather a deterministic one, and relating more to multiplication processes than to additions.

Two well-known sequences, the factorial sequence  $\{N!\} = \{1!, 2!, 3!, 4!, \dots\}$  as well as the self-powered sequence  $\{N^N\} = \{1^1, 2^2, 3^3, 4^4, \dots\}$  are logarithmic in the limit as  $N$  becomes large. Both sequences are considered to be deterministic multiplication processes, yet they differ profoundly from the classic sequential multiplicative processes (such as the generic exponential growth type) since there exists no obvious connecting link between elements in the way of a simple multiplicative factor as in  $X_{N+1} = \text{Factor} * X_N$ .

Prime numbers are not logarithmic, although there is an almost consistent pattern of monotonically decreasing proportions there, favoring low digits. There is no apparent multiplicative process at play here, although the definition of being a prime employs the concept of multiplications, or rather the absence of any multiplicative factor other than 1 and the number itself. Digital proportions fluctuate depending on how many finite primes are chosen. Considering only the first 1000 primes, namely all the primes up to 7919, yields first-digit distribution of:  $\{16\%, 16\%, 13\%, 13\%, 13\%, 14\%, 12\%, 2\%, 1\%\}$ . Considering only the first 3500 primes, namely all the primes up to 32,609, yields a slightly better result:  $\{34.1\%, 32.3\%, 11.3\%, 4.0\%, 3.7\%, 3.9\%, 3.6\%, 3.6\%, 3.6\%\}$ . Considering only the first 6000 primes, namely all the primes up to 59,359, yields yet a different result:  $\{19.9\%, 18.8\%, 18.3\%, 17.8\%, 16.6\%, 2.3\%, 2.1\%, 2.1\%, 2.1\%\}$ . No apparent convergence is seen here in the limit when the number of primes becomes quite large, at least not as far as can be checked with the aid of a simple computer.

Hill and Berger (2007) demonstrated that the root approximation sequence in Newton's Method is logarithmic. The method of finding a root  $X_R$ , namely  $f(X_R) = 0$ , is by the iterative transformation of an initial guessed  $X_G$  value into more exact approximations, using the transformation function for the next root  $X_{N+1} = X_N - f(X_N)/f'(X_N)$  —  $f'(x)$  standing for the derivative of  $f(x)$ . In order to get such a sequence close to the logarithmic two conditions must be met: (i) the initial guess must be quite far from the true root, so that plenty of meaningful iterations are needed, and (ii) to stop as soon as differences between iterated sequence and true root are small, otherwise the sequence would have (almost) all (of) its elements with the same leading digit (that of the root itself).

## DENSITY CURVES AND THEIR LEADING DIGITS DISTRIBUTIONS

---

Approximating probability density function (pdf) for a given data set is done by way of multiplying the vertical scale of its histogram by an adjusting factor so that total area under the curve is unity, while the scale of the x-axis remains unaltered. If the data set is large enough then both pdf and histogram should look alike, except for the height. Leading digits distribution of any data set is uniquely determined by the shape of its histogram/pdf over the relevant x-axis segments. In other words, the distribution of leading digits is determined directly from (I) the shape of the pdf curve above, as well as from (II) the particular range on the x-axis below, indicating what portion of overall data is over which segment on the x-axis where digit  $d$  leads. Figures 4.10 and 4.11 depict eight different histograms whose first leading digits distributions shall now be analyzed.

**Histogram 1** is that of the Uniform(10, 110). Proportion of digit  $d$  leading the first order is calculated simply by the relative widths of all sub-intervals where digit  $d$  is leading as compared with the width of the total range of 100. Since the histogram is uniform, relative width and relative area are equal measures here. Digit 1 leads first digital order on (10, 20) as well as on (100, 110), hence its leadership proportion is  $20/100$ , or 20%. All other digits lead with  $10/100$ , or 10% proportions. Hence first leading digits are  $\{20/100, 10/100, 10/100, 10/100, 10/100, 10/100, 10/100, 10/100, 10/100, 10/100\}$ , or  $\{20\%, 10\%, 10\%, 10\%, 10\%, 10\%, 10\%, 10\%, 10\%, 10\%\}$ . It is noted that second and all higher-orders digit distributions are equal here. For example, digit 0 second-leads on (10, 11); digit 1 second-leads on (11, 12); digit 9 second-leads on (19, 20); and the cycle repeats again, with digit 0 second-leading on (20, 21); digit 1 second-leading on (21, 22); digit 9 second-leading on (29, 30); and so forth. Since all 10 digits obtain equal x-axis segments here, areas (proportions) are also equal for this flat/uniform histogram.

**Histogram 2** is that of the Uniform(10, 100), and its first digit distribution is uniform at  $1/9$  per digit, or 11.1%. This is the naïve intuition we all had

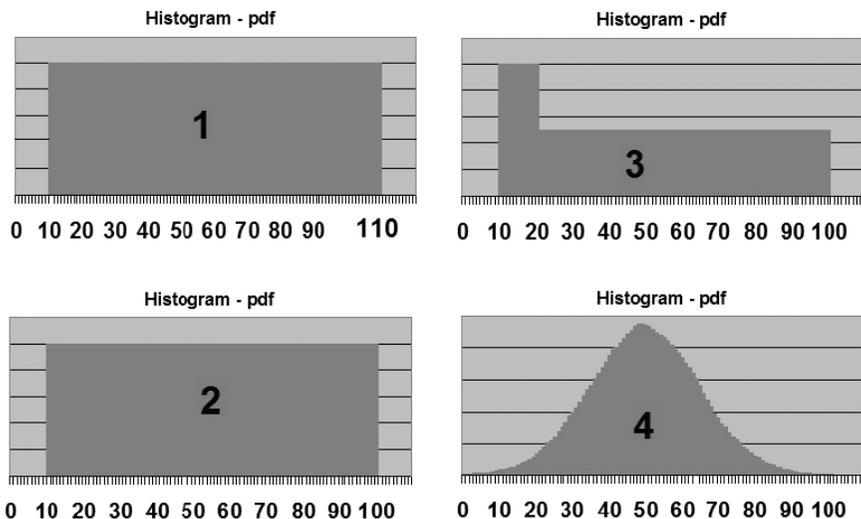


Figure 4.10 Examples of Histograms and their Digit Distributions (1 to 4)

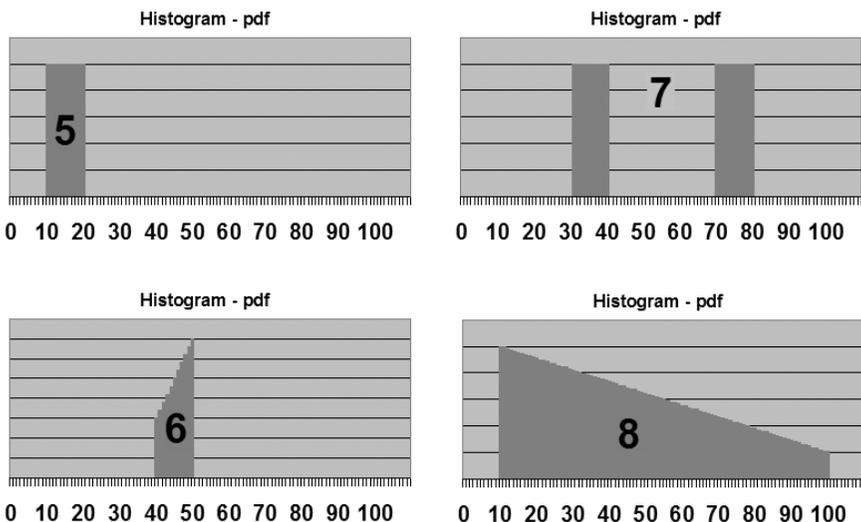


Figure 4.11 Examples of Histograms and their Digit Distributions (5 to 8)

prior to Newcomb and Benford. Note that also second and all higher-order digit distributions are equal here.

**Histogram 3** has the same leading digits distribution as in histogram 1! The sub-interval where digit 1 leads is only (10, 20), and like all other digits it is 10-unit

wide, yet histogram's height is double for digit 1 as compared with the height for all the other digits, thus it has double area, and digital distribution is skewed in favor of digit 1 at: {20%, 10%, 10%, 10%, 10%, 10%, 10%, 10%, 10%}.

Histograms 1 and 3 illustrate a general principle in leading digits, namely the fact that **histogram/pdf uniquely determines leading digits distribution, whereas leading digits distribution does not determine any unique histogram/pdf**. For any given leading digits configuration, there are in principle infinitely many possible histograms/pdf compatible with it. Yet, knowing the leading digits configuration excludes other infinitely many histograms that are incompatible with it. Histograms 1 and 3 also illustrate another concept in leading digits, namely that leadership proportion of any particular digit  $d$  is driven by the confluence of two factors: (I) the relative length on the x-axis over which pdf is defined and where digit  $d$  leads, as well as (II) the relative height of the curve hanging above the segment where digit  $d$  is leading.

**Histogram 4** is that of the Normal(50, 15). Leading digits distribution here is: {2.0%, 7.1%, 16.2%, 24.9%, 24.6%, 15.9%, 7.1%, 1.9%, 0.3%}. This is derived by calculating areas under all relevant nine sub-intervals, (10, 20), (20, 30), (30, 40), ... , (90, 100), and ignoring the extremely tiny and irrelevant fraction that managed to creep below 10 and above 100. For example, the definite integral  $\int_{10}^{20} e^{-\frac{(x-50)^2}{2 \cdot 15^2}} / (15 \cdot \sqrt{2\pi}) dx$  over (10, 20) yields the proportion of digit 1.

**Histogram 5** hands all its first-order digital leadership to digit 1, {100%, 0, 0, 0, 0, 0, 0, 0}. Yet, its second, third, and all higher-order distributions are equally and fairly distributed.

**Histogram 6** hands all its first-order leadership to digit 4. Note that higher-order distributions are all skewed in favor of the high digits here, an inverse configuration!

**Histogram 7** hands half of its first-order leadership to digit 3, and the other half to digit 7, while its second, third, and higher-order distributions are equally and fairly distributed.

**Histogram 8** is a poor attempt at imitating the logarithmic haphazardly. First digits are: {17.6%, 16.0%, 14.4%, 12.7%, 11.1%, 9.5%, 7.9%, 6.2%, 4.6%}. The ratio of the height of the top edge on the left to the height of the lower edge on the right is five. Any other attempt using a straight line that is either steeper or flatter than the particular line shown here — in the hope of improving digital configuration — is due to a decisive failure! The ratio of approximately five seems

to yield approximately the closest one can get to the logarithmic in a linear way. The logarithmic configuration on an interval such as (10, 100) is compatible only with a curved histogram, not a linear one, as shall be shown and discussed in later chapters.

In general, for any pdf curve  $f(x)$ , the probability of digit  $d$  leading first place is given by the expression:  $\int f(x) dx$  over all sub-intervals where digit  $d$  dominates first place. For example, to obtain the probability of digit 7, we need to definite-integrate the probability density function on the sub-intervals:

..., (0.07, 0.08), (0.7, 0.8), (7, 8), (70, 80), (700, 800), (7000, 8000),...

and in general:

$$P(\text{1st digit is } d) = \sum_{N=-\infty}^{+\infty} \int_{d \cdot 10^N}^{(d+1) \cdot 10^N} f(x) dx$$

When working with histograms instead of pdf, it is necessary to divide the sum of all areas where digit  $d$  leads by a denominator expressing the entire area of the histogram, so that the overall proportion of these particular areas within the entire histogram (their probability) is obtained.

## THE CASE OF K/X DISTRIBUTION

---



---

The probability density function  $f(x) = k/x$  plays a fundamental role in Benford's Law, and studying all aspects of this distribution is essential for a complete understanding of the phenomena.

For the probability density function of the form  $f(x) = k/x$  over the interval  $[10^S, 10^{S+G}]$  where  $S$  is any real number,  $G$  is any positive integer, and  $k$  is a constant depending on the values of  $S$  and  $G$ , the sum of all the areas under the curve where digit  $d$  leads is indeed  $\text{LOG}_{10}(1+1/d)$ , namely that  $k/x$  is perfectly logarithmic in the first-order sense. If  $G$  is fairly large, then the requirement that  $G$  must be an integer can be relaxed and digital results are logarithmic (only) in the approximate. Note that the distribution considered here has its long tail to the right abruptly cut at  $10^{S+G}$ , its head abruptly launched at  $10^S$ , and that exponents of 10 representing the two boundaries of the interval differ exactly by an integer, namely  $G$ .

For example,  $k/x$  over the following four intervals is perfectly logarithmic:

$[10^1, 10^{1+1}]$  corresponds to  $[10, 100]$ ,

$[10^2, 10^{2+3}]$  corresponds to  $[100, 100000]$ ,

$[10^{2.58}, 10^{2.58+2}]$  corresponds to  $[380.20, 38020]$ ,

$[10^{1.301}, 10^{1.301+2}]$  corresponds to  $[20, 2000]$ .

For all the above four intervals in the form  $[a, b]$ , the value  $b$  is obtained by simply shifting the decimal point of  $a$  to the right a few places. On the other hand,  $k/x$  over  $[10^{2.5}, 10^{2.5+0.8}]$  corresponding to  $[316.23, 1995.26]$  is not logarithmic in the least, due to that non-integral  $G$  exponent difference of 0.8. [Note: we shall use the notation  $\ln(q)$  or  $\ln q$  to mean the natural logarithm base  $e$  of the number  $q$ ].

Let us prove the above assertion that such  $k/x$  curve is perfectly logarithmic. We first note that the entire area should sum to one, that is  $\int k/x \, dx = 1$  over  $[10^S, 10^{S+G}]$ , therefore  $k[\ln(10^{S+G}) - \ln(10^S)] = 1$ , or  $k[(S+G)\ln 10 - (S)\ln 10] = 1$ , so that  $k*\ln 10*[(S+G) - (S)] = 1$ , and finally  $k*\ln 10*G = 1$ , namely that  $k = 1/[G*\ln 10]$ . Notice that this determination of  $k$  was in total generality,

where  $G$  can assume any value and is not necessarily an integer, and that  $G$  represents the difference in the exponents of the two boundary points spanning the entire interval in question.

Secondly, given a particular pdf of the form  $k/x$  on  $(a, b)$ , we note that for any two sub-intervals within  $(a, b)$  having the same exponent difference, their areas under the curve are identical. Given  $[10^P, 10^{P+R}]$  and  $[10^Q, 10^{Q+R}]$  both contained inside  $(a, b)$ ,  $P$  and  $Q$  being any set of real numbers not necessarily integers, the values of their related constants  $k$  are identical since they belong to the same distribution defined on  $(a, b)$ . The areas under the curve are  $k[\ln(10^{P+R}) - \ln(10^P)]$  and  $k[\ln(10^{Q+R}) - \ln(10^Q)]$  respectively, or simply  $k[(P+R)\ln 10 - (P)\ln 10]$  and  $k[(Q+R)\ln 10 - (Q)\ln 10]$ . Further simplifying we get:  $k*\ln(10)*[(P+R) - (P)]$  and  $k*\ln(10)*[(Q+R) - (Q)]$ , which yields  $k*\ln 10*R$  as the same area for each sub-interval. If the whole interval  $(a, b)$  is expressed as  $[10^S, 10^{S+G}]$  so that  $k = 1/[G*\ln 10]$  then area for each is simply  $R/G$ , namely the ratio of exponent difference of the sub-interval to the exponent difference of the entire range.

For example, for  $k/x$  defined on  $(1, 10000)$ , the sub-intervals  $[1, 10]$  and  $[100, 1000]$  have equal areas, since their exponent differences are identical, namely 1. While  $[1, 10]$  is narrower on the  $x$ -axis, pdf value hanging above is high. On the other hand  $[100, 1000]$  is extremely long on the  $x$ -axis in comparison, but its pdf above is quite low. This trade-off exactly cancels out the effect of each factor so that areas end up the same.

To set the stage for the proof, we break  $S$  into  $N + f$ , an integral  $N$  component and a fractional  $f$  component, and represent the entire interval in question as  $[10^{N+f}, 10^{N+f+G}]$ , where  $N$  is zero or some positive integer,  $f$  is some possible fractional part, namely  $0 \leq f < 1$ , and  $G$  is a positive integer representing the integral difference in exponents.

We initially let  $f = 0$ , a restriction that would be relaxed later. Considering any digit  $D$ , the probability that  $D$  leads is given by the area under  $k/x$  on the following intervals:

$$\{[D10^N, (D+1)10^N], [D10^{N+1}, (D+1)10^{N+1}], [D10^{N+2}, (D+1)10^{N+2}], \dots \mathbf{G \text{ times}} \dots, [D10^{(N+G-1)}, (D+1)10^{(N+G-1)}]\}.$$

This is so because it is on these intervals and these intervals alone that  $D$  leads. Calculating the various definite integrals we obtain:

$$k[\ln((D+1)10^N) - \ln(D10^N)] + k[\ln((D+1)10^{N+1}) - \ln(D10^{N+1})] + \dots \mathbf{G \text{ times}} \dots + k[\ln((D+1)10^{N+G-1}) - \ln(D10^{N+G-1})] \text{ or:}$$

$$\begin{aligned}
 &k[\ln(D+1) + N*\ln(10) - \ln(D) - N*\ln(10)] + \\
 &k[\ln(D+1) + (N+1)*\ln(10) - \ln(D) - (N+1)*\ln(10)] + \dots \mathbf{G \text{ times}} \dots + \\
 &k[\ln(D+1) + (N+G-1)*\ln(10) - \ln(D) - (N+G-1)*\ln(10)]
 \end{aligned}$$

Canceling out like terms (not involving D) we are left with:

$$\begin{aligned}
 &k[\ln(D+1) - \ln(D)] + k[\ln(D+1) - \ln(D)] + \dots \mathbf{(G \text{ times})} \dots + k[\ln(D+1) - \ln(D)] = \\
 &k*\ln[(D+1)/(D)] + k*\ln[(D+1)/(D)] + \dots \mathbf{(G \text{ times})} \dots + k*\ln[(D+1)/(D)] = \\
 &G*k*\ln[(D+1)/(D)].
 \end{aligned}$$

Substituting here the expression for k above we obtain:

$$G*(1/[G*\ln10])*ln[(D+1)/(D)] = ln[(D+1)/D]/ln10.$$

This expression uses the natural logarithm base e. Applying the logarithmic identity  $LOG_A X = LOG_B X / LOG_B A$  (twice) to convert this ratio to the common logarithm base 10 yields:

$$\begin{aligned}
 &LOG_{10}[(D+1)/D] / LOG_{10}[e] \ / \ (LOG_{10}[10] / LOG_{10}[e]) \\
 &LOG_{10}[(D+1)/D] / LOG_{10}[e] \ / \ (1 / LOG_{10}[e]) \\
 &LOG_{10}[(D+1)/D] \\
 &LOG_{10}[1+1/D]!
 \end{aligned}$$

Let us now show why  $f \neq 0$  should not yield any different result. As shown above k depends solely on the difference between the exponents [as can be seen from its expression  $k = 1 / (G*\ln10)$ ], hence letting  $f \neq 0$  (a slight shift to the right from the  $f = 0$  situation) does not alter that exponent difference, and k is still of the same value (whether f equals to zero or not). Thus the form of the function  $k/x$  is not altered either, only its defined range is being shifted. Now, since areas under the curve are identical for any two sub-intervals of the same exponent difference as was shown earlier, it follows that any non-zero f that requires an additional area to the **right** of the intervals  $\{[D10^N, (D+1)10^N], [D10^{N+1}, (D+1)10^{N+1}], \dots, [D10^{(N+G-1)}, (D+1)10^{(N+G-1)}]\}$  would then also require an identical subtraction on the **left** side of those intervals. This completes the proof.

An integral difference G in the exponents of the interval  $[10^S, 10^{S+G}]$  was required for exact logarithmic behavior, yet if G is fairly large (depending on precision required), then the requirement can be relaxed without effecting much the above proof. This is so because the value of the constant k is  $1/[G*\ln10]$ , and as such it is inversely proportional to G, hence each term in the earlier expression  $k*\ln[(D+1)/(D)] + k*\ln[(D+1)/(D)] + \dots \mathbf{(G \text{ times})} \dots + k*\ln[(D+1)/(D)]$  becomes quite marginal in the grand scheme of things for large values of G (and

the implied small values of  $k$ ). Put another way, since  $G$  signifies approximately the number of sub-intervals standing between IPOT points, it follows that when a large value of  $G$  comes with a fractional part, distorted digital proportions on the first and the last sub-intervals between IPOT constitute small portions of overall data, and their digital influence is quite insignificant.

As an example, for  $k/x$  defined over  $(10, 3000000000)$ , namely on the interval  $[10^1, 10^{9.48}]$ ,  $G = 8.48$ , and  $k = 1/[G*\ln 10] = 1/[8.48*\ln 10] = 0.0512$ . Although  $G$  is not an integer here, yet since it's quite large, first digits distribution comes at  $\{31.5\%, 19.2\%, 12.0\%, 9.2\%, 7.2\%, 6.3\%, 5.3\%, 4.8\%, 4.5\%\}$ , and this result is very near the logarithmic. Digits 1 and 2 got an extra portion over and above the logarithmic on the marginal range of  $[1000000000, 3000000000]$  on the right, taking a little bit of leadership from all the other digits, but this is still a minor portion in the grand scheme of things.

Let us illustrate digital proportions for the  $k/x$  case defined over  $(10, 100)$ , that is, over  $[10^1, 10^2]$ . Here,  $G = 1$ , and  $k = 1/[G*\ln 10] = 1/[1*\ln 10] = 1/[2.302585] = 0.4342945$ . Figure 4.12 depicts this particular  $0.4342945/x$  density curve defined over  $(10, 100)$ .

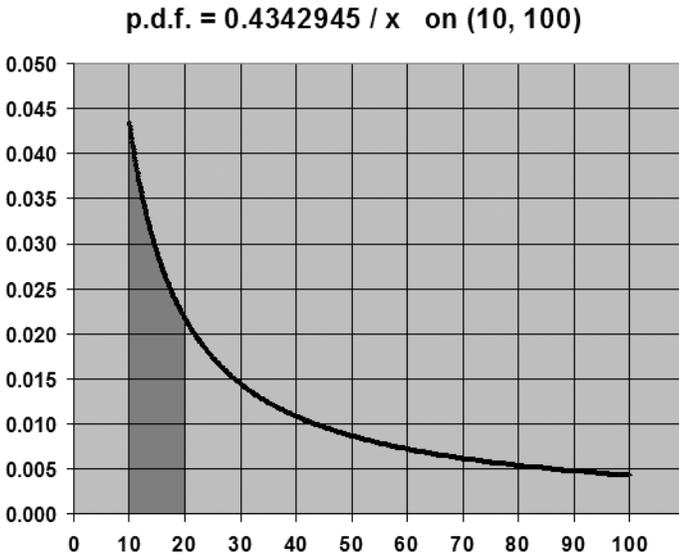
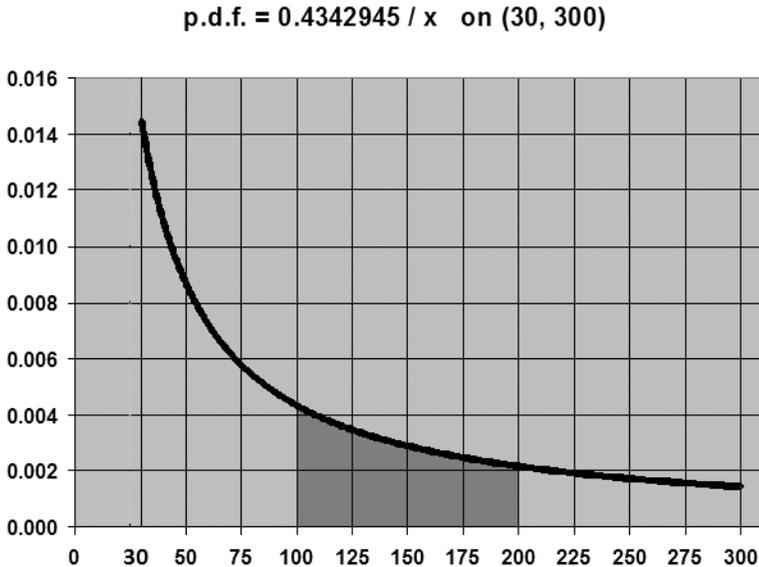


Figure 4.12 Perfectly Logarithmic Distribution:  $0.4342945/x$  Defined on  $(10, 100)$

The reader is invited to easily calculate area under this curve for any digit  $D$  to determine by how much it is leading. For example, digit 1 leads only on the sub-interval  $[10, 20)$  where area there is roughly  $(20-10)*\text{average height} = (10)*[(0.043 + 0.022)/2] = 10*0.0325 = 0.325$  or 32.5%. Although this should be a bit less as curve is concaved inwards bumping slightly into the line of the average, so true result is the value of 30.1% as in Benford's Law. Digit 2 leads only on the sub-interval  $[20, 30)$  where area there is approximately  $(30-20)*\text{average height} = (10)*[(0.022 + 0.014)/2] = 10*0.018 = 0.180$  or 18.0%, which after a bit of subtraction due to the curve concaving inward, it is exactly 17.6% as in Benford's Law. This easy visualization here occurs only because  $k/x$  is defined exactly between two adjacent (consecutive) integral powers of ten, like 10&100, 1&10, and so forth.

Another example of a perfectly logarithmic density curve is given by  $k/x$  distribution defined over  $(30, 300)$ , that is, over  $[10^{1.47712}, 10^{2.47712}]$ , hence  $G = 1$ , and  $k = 1/[G*\ln 10] = 1/[1*\ln 10] = 1/[2.302585] = 0.4342945$ . Figure 4.13 depicts this particular  $0.4342945/x$  density curve defined over  $(30, 300)$ .

As compared with  $k/x$  over  $(10, 100)$ , here data starts 'artificially' at 30, yet it's logarithmic just the same because it ends 'smartly' exactly at 300. Digit 1,



**Figure 4.13** Perfectly Logarithmic Distribution:  $0.4342945/x$  Defined on  $(30, 300)$

which supposedly lost leadership completely on  $[10, 20)$ , has regained exactly that portion, and now leads exclusively on  $[100, 200)$ , which has a longer span on the x-axis than that of  $[10, 20)$  but comes with a lower curve, a perfect trade-off. A very rough estimate for digit 1 area of leadership is  $(200-100)*\text{average height} = (100)*[(0.004 + 0.002)/2] = 100* 0.003 = 0.30$  or 30.0%, quite near the true value of 30.1% in Benford’s Law.

An example of a **non-logarithmic**  $k/x$  distribution is given by  $k/x$  defined over  $(10, 200)$ , that is, over  $[10^1, 10^{2.30103}]$ , hence  $G = 1.30103$  (non-integral), and  $k = 1/[G*\ln 10] = 1/[1.30103*\ln 10] = 1/[2.9957] = 0.3338082$ . Figure 4.14 depicts this particular  $0.3338082/x$  density curve defined over  $(10, 200)$ .

Here digit 1 is forcefully usurping power. It leads even more than it is normally supposed to in logarithmic distributions. Digit 1 leads on  $[10, 20)$  as well as on  $[100, 200)$ , way above its 30.1% supposed portion, taking leadership from all other digits. The first-digit distribution here is  $\{45.7\%, 13.6\%, 9.6\%, 7.1\%, 6.4\%, 5.1\%, 4.4\%, 4.3\%, 3.8\%\}$ . Upper limit here should have stopped at the 100 mark in order to stay logarithmic and to prevent that digital coup d’etat from occurring, but it overshot it. On the other hand, as discussed in the proof above, for  $k/x$  defined over an interval composed of a large number of IPOT numbers,

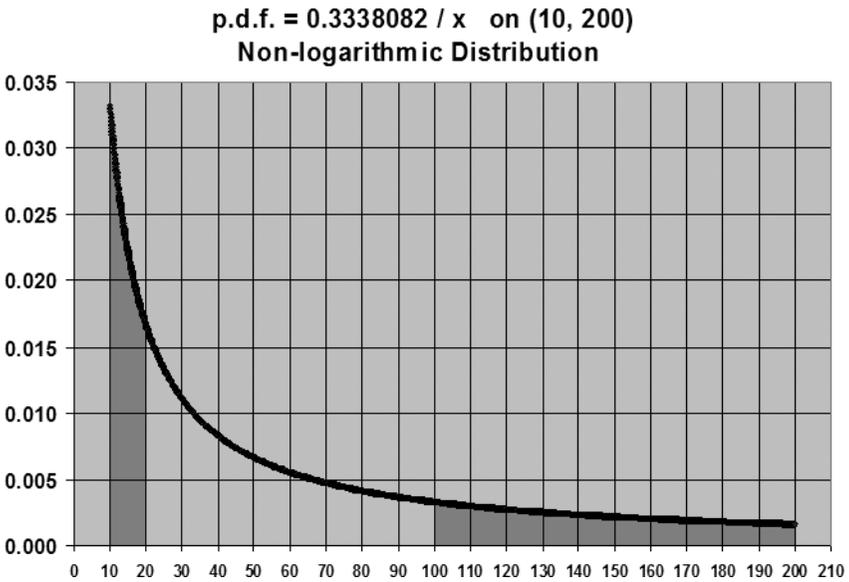


Figure 4.14 A Non-Logarithmic Distribution:  $0.3338082/x$  Defined On  $(10, 200)$

where  $G$  is quite large, overshooting supposedly ‘proper interval’, where  $G$  is not an integral value anymore, does not disrupt digit configuration much from their logarithmic behavior, because that extra area becomes marginal in the grand scheme of things.

As for another example,  $k/x$  defined over  $(1, 2,000,000,000,000)$  is nearly logarithmic. That ‘extra’ huge sub-interval  $(1,000,000,000,000, 2,000,000,000,000)$  should not falsely alarm us as significant, because the portion of overall area (data) there is not very large, and in fact it’s exactly equal to the portion on say  $(10, 20)$  having identical exponent difference, as was discussed earlier (because  $k/x$  curve on 1 to 2 trillion is ‘very low’.)

The distribution  $k/x$  defined over  $(10, 90)$  isn’t logarithmic either, as digit 9 is totally deprived of leadership. Another extreme example is the distribution  $k/x$  defined over  $(30, 100)$  which isn’t logarithmic because digits 1 and 2 are totally deprived of leadership.

Let us consider  $k/x$  defined over an extremely large interval  $(a, b)$ , say  $a = 10^1$  and  $b = 10^{15}$ . **The conceptual assertion that  $k/x$  on  $(a, b)$  is ‘continuously logarithmic’** can be justified or illustrated by sliding to the right or to the left a flexible imaginary lens, focusing solely on the much smaller sub-interval  $[10^S, 10^{S+1}]$  contained anywhere inside  $(a, b)$ , and letting  $S$  vary. Here, exponent difference is always 1 no matter what value  $S$  assumes. The width of the lens (on the  $x$ -axis) widens as we slide to the right (but height falls), and shrinks as we slide to the left (but height rises), while data generated from it in accordance with local density is always exactly and perfectly logarithmic! On the other hand, for a non-logarithmic  $k/x$  sub-interval defined over  $[10^S, 10^{S+1.8}]$  where exponent difference is a non-integral 1.8, displacements affects digital distribution a great deal.

## UNIFORM MANTISSA, VARIED SIGNIFICAND, AND THE GENERAL LAW

---

In the first paper ever to acknowledge the phenomena, Newcomb asserted that “the law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable”. This alternative vista of numbers and significant digits, viewing them via the sharply focused lens of the mantissa, significantly thrusts understanding of the phenomena forward, leading us to discover many of its most important aspects and to facilitate an alternative explanation. The singular form of the term is ‘mantissa’, and the plural form is ‘mantissae’. Admittedly, at first the concept of the mantissa appears somewhat alien, being at odds with our intuitive idea of numbers. In the literature, the mantissa at times is said to be simply **‘the fractional part of the log’**, although this definition is too simplistic since it is not true for numbers less than 1. The definition of the mantissa of any number  $X$  (assumed to be positive) is the unique solution to  $X = 10^W * 10^{\text{mantissa}}$ , where  $W$  — called the ‘characteristic’ — is the result of rounding down  $\log(X)$  to the nearest integer [the largest integer less than or equal to  $\log(X)$ ], or equivalently the first integer to the left on the real number line  $(-\infty, +\infty)$  of the log-axis, and regardless whether  $\log(X)$  is positive or negative. For example, for  $X = 750$ ,  $\log$  is 2.875, the characteristic is 2, and mantissa is 0.875. For  $X = 0.077$ ,  $\log$  value is  $-1.114$ , the characteristic is  $-2$ , and mantissa is 0.886. Note with caution to avoid confusion that when  $\log(X)$  is negative, the characteristic is some negative integer as well, and that rounding down the log actually implies a larger absolute value for the characteristic. **For any number  $X$  equal to or bigger than 1, mantissa is simply the fractional part of  $\log(X)$** , and the characteristic is the integral part of  $\log(X)$ , [ $X \geq 1$  implies that  $\log(X)$  is positive or zero]. For any number  $X$  less than 1, mantissa is simply 1 minus (or the one-complement of) the fractional part of the absolute value of  $\log(X)$ , [ $X < 1$  implies that  $\log(X)$  is negative]. By convention, when  $\log$  is exactly an integer, mantissa is thought to be zero, and the characteristic is just that integral log value. Therefore mantissa  $\in [0, 1)$ . It should be noted with caution that for  $X = 0.01259$  for example, being a value less

than 1, its log is  $-1.900$ , characteristic is  $-2$ , and mantissa is the distance on the log-axis between the log and the characteristic, namely  $0.100$ , and not simply the fractional part of the log! Therefore, **a preferred view of the mantissa of X is the distance on the log-axis between  $\log(X)$  and the integer immediately to the left of it.**

For negative values, that is when  $X < 0$  and  $\log(X)$  is undefined, the above discussion and definitions simply applies to  $|X|$  substituted for  $X$ . Conceptually, the mantissa can be thought of as that unique value representing the relative location of the number in question in between its adjacent IPOT values immediately to the left and to the right of it.

The **significant** of  $X$  (assumed to be positive) is defined as the unique solution to  $X = 10^W * S$ , where  $W$  is the result of rounding down  $\log(X)$  to the nearest integer (the 'characteristic'). In other words, significant is simply the way the number is written and expressed (using digits) disregarding the decimal point by always placing it second from the left for any value. For example, if  $X = 175.83$ , significant is  $1.7583$ . Hence significant is that first part of scientific notation with the exponential part totally ignored. Calculation of  $\log(X)$  is actually totally unnecessary for finding significant of  $X$ . Note that significant  $\in [1, 10)$ . For negative values, that is when  $X < 0$  and  $\log(X)$  is undefined, the above definition simply applies to  $|X|$  substituted for  $X$ . Figure 4.15 gives some concrete examples of the above definitions and concepts.

Number	LOG	Mantissa	Significant	1st Digit
8	0.903	0.903	8.0000	8
80	1.903	0.903	8.0000	8
800	2.903	0.903	8.0000	8
230.65	2.363	0.363	2.3065	2
0.00022	-3.658	0.342	2.2000	2
0.0818	-1.087	0.913	8.1800	8
56	1.748	0.748	5.6000	5
34566	4.539	0.539	3.4566	3
7	0.845	0.845	7.0000	7
3	0.477	0.477	3.0000	3
10	1	0	1.0000	1
100	2	0	1.0000	1
1000	3	0	1.0000	1
0.01	-2	0	1.0000	1

**Figure 4.15** Examples of Numbers, Their Log, Mantissa, Significant, and First Digit

From the two definitions above,  $X = 10^W * 10^{\text{mantissa}}$  and  $X = 10^W * S$ , it follows that the relationships between mantissa and significand are:

$$\text{significand} = 10^{\text{mantissa}}$$

$$\text{mantissa} = \text{LOG}_{10}(\text{significand})$$

Following the curve of the mantissa and significand (as a function of  $x$ ) along their journeys throughout the  $x$ -axis, we notice a perfect repetition after each IPOT number. The two equations below, which follows directly from the definitions, as well as the graph in Fig. 4.16 for the significand defined over (10, 1000), illustrate this principle:

$$\begin{aligned} \text{mantissa of } (10^N * X) &= \text{mantissa of } (X) & [N \text{ being any integer}] \\ \text{significand of } (10^N * X) &= \text{significand of } (X) & [N \text{ being any integer}] \end{aligned}$$

Consider for example the two numbers 4.782 and 4782. Both are with the same leading digits (all orders considered, that is  $D_1 = 4$ ,  $D_2 = 7$ ,  $D_3 = 8$ , and  $D_4 = 2$ ), of the same mantissa, and of the same significand. The difference in their logarithms is exactly an integer, 3 in this case, and which can be completely ignored as far as leading digits, mantissa, and significand are concerned. This example illustrates the common thread going through these three concepts. **In other words, leading digits (all orders included), significand, and mantissa, are simply three different ways to express the same concept, and knowing the value of any one of the three uniquely points to the values of the other two.**

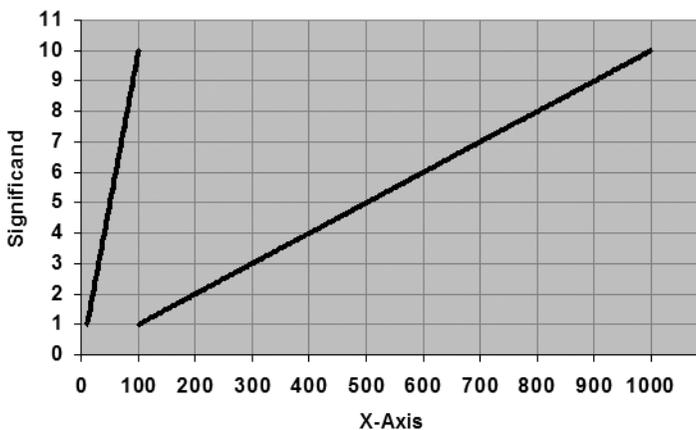
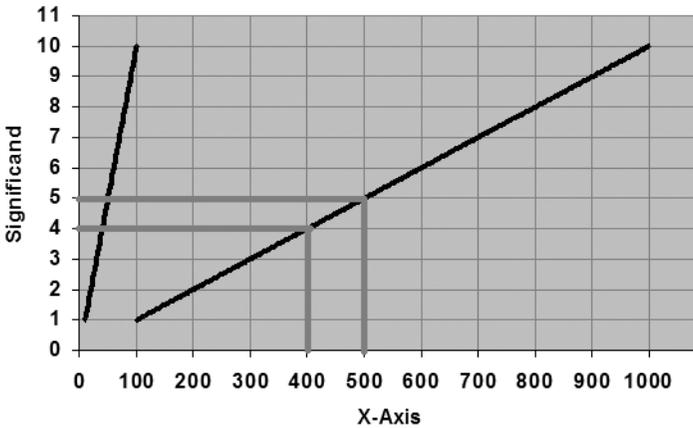


Figure 4.16 The Significand as a Function of X on (10, 1000)



**Figure 4.17** Visualizing the Correspondence between Significand and Leading Digits

Let us recap: a given significand uniquely determines all orders of leading digits, and vice versa. The same applies to mantissa which has a one-to-one correspondence with the significand. A slightly better insight is gained by the visualization given in the graph of Fig. 4.17, showing how the sub-region of (400, 500) on the x-axis corresponds with the y-axis (4, 5) values of the significand. The same types of gray lines in the graph should be drawn from the x-axis sub-region of (40, 50) to the corresponding identical significand segment of (4, 5), and so forth for any other sub-region ( $4 \cdot 10^{\text{INTEGER}}$ ,  $5 \cdot 10^{\text{INTEGER}}$ ).

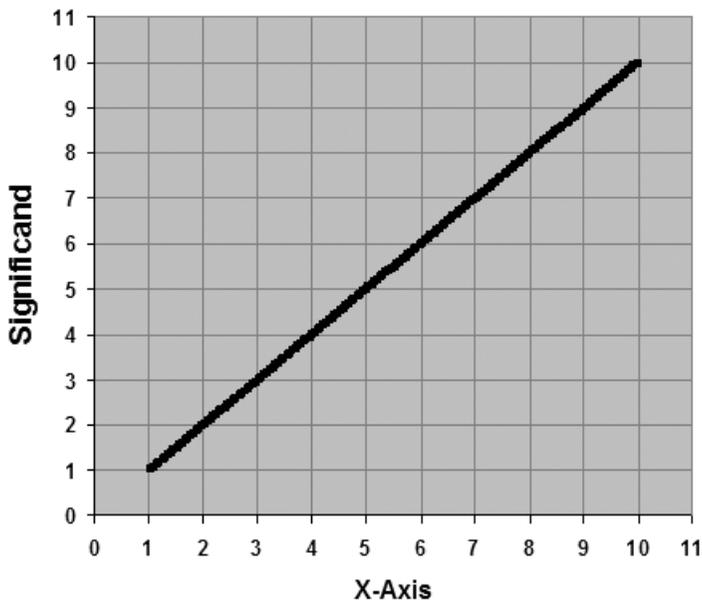
To emphasize the point that significand is nothing but the rewriting of the number in question with all its digits and digital orders intact while ignoring the decimal point, the chart in Fig. 4.18 is added, showing the significand function on the (1, 10) range, in which case, there is nothing actually to modify in the numbers!

We now have the tools to properly define Benford's Law in its general form, detailing exactly how digits occur in total generality. Remarkably, the general statement reveals in one decisive stroke all there is to know about digital distributions, be it the first order, higher orders, or the relationships and dependencies between the orders themselves.

#### **THE GENERAL FORM OF BENFORD'S LAW:**

$$\text{Probability}(\text{significand} \leq S_0) = \text{LOG}_{10}(S_0)$$

$S_0$  is an element of [1, 10).



**Figure 4.18** Significand as a Function of X on (1, 10) is Simply X Itself

The elegance, simplicity, and compactness of the general law are quite striking. It is exhilarating to see that such a concise expression covers the entire phenomena! To further generalize this statement, other bases may be used as in:  $\text{Probability}(\text{significand} \leq S_0) = \text{LOG}_{\text{BASE}}(S_0)$ .

Note that an alternative expression of the general form of Benford's Law could be written, utilizing the relationship  $\text{mantissa} = \text{LOG}_{10}(\text{significand})$ , thus yielding:  $\text{Probability}(\text{significand} \leq S_0) = \text{mantissa}(S_0)$ .

The General Law, which is stated in terms of significand, specifies exactly the probability of any combination of digits. This is so because of the one-to-one relationship between significand and leading digits. For example, the probability that the first digit would be 4 and the second digit would be 7 is given by the General Law statement  $\text{P}(4.7 \leq \text{Significand} < 4.8) = \log(4.8) - \log(4.7) = 0.681 - 0.672 = 0.009$ , or 0.9%. Previously we applied the equivalent expression  $\text{LOG}(1+1/47) = 0.009$ . Also,  $\text{P}(\text{2nd digits are 3, while first digits attaining any configuration}) = \text{P}(1.3 \leq S < 1.4) + \text{P}(2.3 \leq S < 2.4) + \dots + \text{P}(9.3 \leq S < 9.4) = \log(1.4) - \log(1.3) + \log(2.4) - \log(2.3) + \dots + \log(9.4) - \log(9.3) = 0.1043$ , or 10.43%. In fact this is the calculation of the unconditional second-order probability for digit 3. Interestingly and perhaps unexpectedly, the General Law

implies that the probabilities of the orders are dependent on each other, that is, probabilities positively correlate with each other to some degree. For example, the probability of second-order digits varies depending on what happened to first digit, and vice versa. More specifically, calculations show that there is 0.109 unconditional probability that the second digit is 2 (low digit itself). Yet, the conditional probability that second digit is 2 given that the first digit happened to be 1 (low digit) is 0.115 (higher than 0.109), while the conditional probability that the second digit is 2 given that the first digit happened to be 9 (high digit) is 0.103 (lower than 0.109). In summary, the General Law shows that a low preceding digit implies slightly higher probability for another low digit to follow (as compared with the unconditional probability of that low digit), while a high preceding digit implies slightly lower probability for the same low digit.

**The general form of Benford's Law implies  $\text{LOG}_{10}(1+1/d)$ :** First to note is that digit  $d$  (1 to 9) leads first whenever significand is in  $[d, d+1)$ . In other words, viewing digit  $d$  as a significand, namely as  $d.0000000000$ , and not merely as an integral digit. For example, digit 3 leads first whenever significand is within the interval  $[3.00000, 4.00000)$ . Hence:  $P(d \text{ is first}) = P(d \leq \text{significand} < d + 1)$ , or:  $P(\text{significand} < d + 1) - P(\text{significand} < d)$ , which is translated via the general form as:  $\text{LOG}_{10}(d+1) - \text{LOG}_{10}(d) = \text{LOG}_{10}[(d+1)/(d)] = \text{LOG}_{10}[1+1/d]$ .

**The general form of Benford's Law implies uniformity of mantissa:**  $P(\text{significand} < S_0) = \text{LOG}_{10}(S_0)$  pertains to some particular significand  $S_0 \in [1, 10)$ , but it can also be thought of as pertaining to some particular mantissa  $M_0 \in [0, 1)$  such that  $M_0 \equiv \text{LOG}_{10}(S_0)$ . This is so since mantissa and significand have a one-to-one correspondence in the form  $\text{mantissa} = \text{LOG}_{10}(\text{significand})$ , and so the probability of having all types of significand less than  $S_0$  is equal to the probability of having all types of mantissa less than  $M_0$ . The General Law can then be written as:

$$P(\text{mantissa} < M_0) = \text{LOG}_{10}(S_0), \text{ or } \mathbf{P(\text{mantissa} < M_0) = M_0.}$$

In other words, the cumulative probability of mantissa is the mantissa itself. Writing  $M$  as a random mantissa variable and incorporating the above result yields Distribution Function  $F(M) = M$ . Differentiating with respect to  $M$  yields Probability Density Function  $f(M) = 1$ , namely mantissa with a uniform density of height of 1 over the area  $[0, 1)$ . Since all of the above could easily be done in reverse, this also proves that: **Uniformity of mantissa implies the general form of Benford's Law.**

There are nine compartments within  $[0, 1)$  mantissa, each pointing to a unique **first** digit leading, and whose probabilities are in direct proportion to their width on the M-axis. The nine compartments in question are: **[0, 0.301]**, **[0.301, 0.477]**, **[0.477, 0.602]**, **[0.602, 0.699]**, **[0.699, 0.778]**, **[0.778, 0.845]**, **[0.845, 0.903]**, **[0.903, 0.954]**, and **[0.954, 1.000]**. Such special case occurs when data falls on  $(1, 10)$  following the  $k/x$  density curve.

This sequence of fractions above is the cumulative percentages of  $\text{LOG}_{10}(1+1/d)$ . For example, if data is logarithmic then digit 2 leads first order whenever mantissa  $M$  satisfies  $\text{LOG}(2) \leq M < \text{LOG}(3)$ , namely whenever  $0.301 \leq M < 0.477$ . Hence probability of digit 2 leading first order is 0.176 (simply 0.477 minus 0.301), and its mantissa is confined within the second compartment  $[0.301, 0.477]$ . Since height of  $\text{pdf}(M)$  is constant at 1, areas (representing probabilities) can be calculated by simply reading the width on the M-axis.

Figure 4.19 depicts these nine compartments of mantissa and their corresponding first-order leading digit. The notation 'LOG' is also incorporated into the M-axis for the special case where log of data happened to be uniformly distributed on  $[0, 1)$ , in which case the terms mantissa and log are interchangeable (i.e. whenever fractional part of the log is also log itself). Such special case occurs when data falls on  $(1, 10)$  following the  $k/x$  density curve.

Imagine data having its mantissa distributed over  $[0, 1)$  in a non-uniform non-linear way, with hills and valleys and curves, and even discontinuities, yet having correct proportions in those nine compartments above, in the sense that overall area between 0 and 0.301 is 0.301, and overall area between 0.301 and 0.477 is 0.176, and so forth. In this case  $\text{LOG}_{10}[1+1/d]$  for the first digits law is being strictly observed, but laws for higher orders may be off and the General Law not followed. Smoothness of data typically precludes such occurrences, especially with very large data sets. Figure 4.20 demonstrates one such extreme case having this dichotomy between the first order and the rest of all the higher orders, where continuity is severely violated. The chart in Fig. 4.20 stands for the pfd of the data itself, not for the mantissa. It depicts data falling strictly over the interval  $(10, 100)$ . The smooth gray line for comparison is of the logarithmic density curve  $0.4343/x$  also defined on  $(10, 100)$ . All higher orders here come with equal probability of 10%, and there are no dependencies or correlations in probabilities between the orders. Needless to say, this rarely occurs (if ever) in real-life data sets. Finally, in anticipation of the topics in the next several chapters, the term '**related log**' of data is introduced, referring to the data set generated when each number in the raw data set itself is transformed to its LOG value.

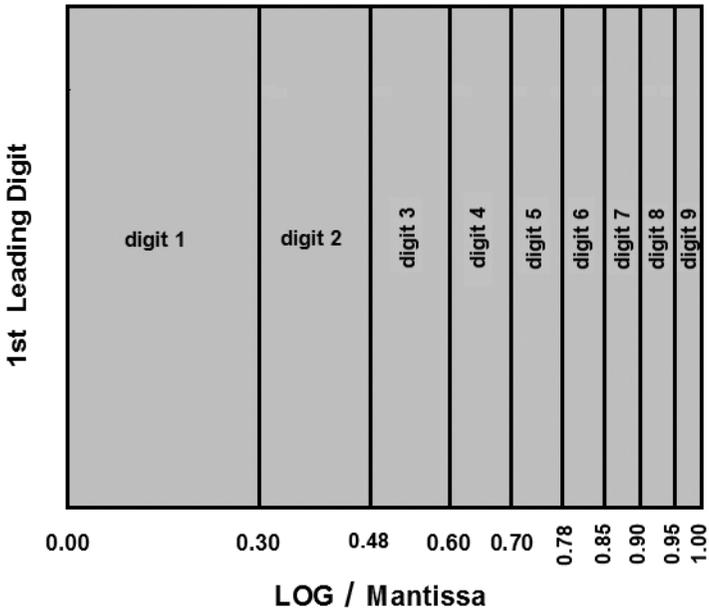


Figure 4.19 Nine Mantissa Compartments, Each Corresponding to a Unique First Digit

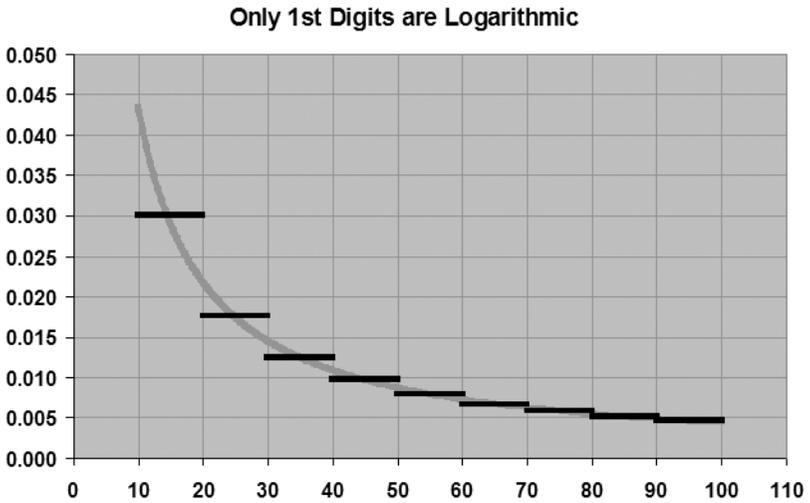


Figure 4.20 Data (shown as black) Being Logarithmic Solely in First Order Sense

## UNIQUENESS OF K/X DISTRIBUTION

---



---

Let us now turn our attention to the most obvious way of obtaining uniformity of mantissa, namely the special case of uniformity of log of data itself with its range on the log-axis measuring exactly some integral value. For example, assume related log of data is uniformly distributed on (7.0, 9.0). Resultant overall mantissa is then calculated by simply aggregating the two mantissa areas belonging to (7.0, 8.0) and (8.0, 9.0), yielding uniform mantissa as well. The same result is gotten for related log uniformly distributed on (4.3, 8.3). On the other hand, if related log of data is uniformly distributed on (5.0, 6.5), resultant overall mantissa is not uniform at all, and digits 1 and 2 lead strongly, way above their normal logarithmic advantage.

Assume  $X$  has pdf  $f(x) = k/x$  over the interval  $[10^S, 10^G]$ ,  $G > S$ , with both  $S$  and  $G$  being any number, integer or non-integer. The Distribution Function Technique can be employed to find the pdf of the transformation  $Y = \text{LOG}_{10}X$ . The distribution function of  $Y$ ,  $F(y)$ , is then  $P[Y < y, \text{ for } S \leq y \leq G] = P[\text{LOG}(X) < y, \text{ for } S \leq y \leq G]$ . Taking 10 to the power of both sides of the inequality we get:  $F(y) = P[10^{\text{LOG}(X)} < 10^y, \text{ for } S \leq y \leq G]$ , that is:

$F(y) = P[X < 10^y, \text{ for } S \leq y \leq G] = \int k/x \, dx$  over the interval  $(10^S, 10^y)$ , so that  $F(y) = k[\ln(10^y) - \ln(10^S)] = k[(y - S) \cdot \ln(10)]$ . Since  $k = 1/(I \cdot \ln(10))$  where  $I$  represents the difference in the exponents of the two boundaries spanning the entire interval length of  $x$ , as was demonstrated earlier regarding  $k/x$ , we finally obtain:

$F(y) = [1/([G - S] \cdot \ln(10))] \cdot [(y - S) \cdot \ln(10)] = (y - S)/(G - S)$ . Differentiating with respect to  $y$  and employing the relationship  $f(y) = dF(y)/dy$ , we finally get: pdf( $Y$ ) =  $1/(G - S)$ , namely a constant probability density function of the form  $1/(\text{exponent difference})$ . Hence:

**Proposition I:** If pdf of  $X$  is of the form  $f(x) = k/x$  over the interval  $[10^S, 10^G]$ ,  $S$  and  $G$  being any real number, then its related log distribution, namely  $Y = \text{LOG}_{10}X$  over  $[S, G]$ , is uniformly distributed with pdf of the form  $f(y) = 1/(G - S)$ .

(No consideration is given here to leading digits).

**Corollary I:** If pdf of X is of the form  $f(x) = k/x$  over the interval  $[10^S, 10^{S+N}]$ , S being any real number, and N — the difference in exponents — is exactly an integer, then X is logarithmic according to the General Law.

This follows directly from Proposition I implying that related log of X is uniformly distributed, while difference in exponents is assumed to be of an integral value, rendering mantissa of X uniformly distributed. As an example, imagine X distributed as  $k/x$  over  $[10^{3.88}, 10^{6.88}]$  with exponent difference of 3. Here log of X is uniformly distributed on  $[3.88, 6.88]$ . Clearly mantissa is uniformly distributed as well; we simply exchange the margin on the left before the integral value of 4 on  $(3.88, 4)$ , with the right margin filling in  $(6.88, 7)$  space, to obtain the wholesome interval  $(4, 7)$  with its uniformity intact.

In the earlier chapter regarding the case of  $k/x$  distribution, it was demonstrated that if the interval over which  $k/x$  is defined is of an integral exponent difference then its first leading digits are logarithmically distributed as in  $LOG[1+1/d]$ . Here we go one step further asserting that an integral exponent difference in the interval for  $k/x$  implies the logarithmic distribution according to the General Law with all higher orders considered.

Assume Y is uniformly distributed over  $[R, S]$ ,  $f(y) = 1/(S - R)$ , the difference  $S - R$  being any number. The Distribution Function Technique can be employed to find the pdf of the transformation  $X = 10^Y$  over the interval  $[10^R, 10^S]$ .

The distribution function of X,  $F(x)$ , is then  $P[X < x, \text{ for } 10^R \leq x \leq 10^S] = P[10^Y < x, \text{ for } 10^R \leq x \leq 10^S]$ .

Taking log base 10 to both sides of the inequality we get:

$F(x) = P[LOG(10^Y) < LOG(x), \text{ for } 10^R \leq x \leq 10^S]$ , that is:

$F(x) = P[Y < LOG(x), \text{ for } 10^R \leq x \leq 10^S] =$

$\int 1/(S - R) dy \text{ over the interval } [R, LOG(x)]$ ,

so that  $F(x) = [LOG(x) - R] * [1/(S - R)]$ . Differentiating with respect to x and employing the relationship  $f(x) = dF(x)/dx$ , we finally get:

$pdf(X) = [1/\ln(10)] * [1/(S - R)] * [1/x]$  distributed over  $[10^R, 10^S]$ . Hence:

**Proposition II:** If Y is uniformly distributed over  $[R, S]$ , with the length  $S - R$  being any real number, then  $X = 10^Y$  over  $[10^R, 10^S]$  is distributed with pdf of the form  $f(x) = (1/[S - R] * \ln(10))] * (1/x)$ .

(No consideration is given here to leading digits).

**Corollary II:** If related log of X is uniformly distributed over  $[R, S]$  with the length  $S - R$  being an integer, then X is logarithmic according to the General Law when considered over its entire range of  $[10^R, 10^S]$ . Moreover, X is then necessarily of the form  $k/x$ .

In other words, if the difference in exponents of the extreme points of X is of an integral value, and if LOG of X is uniformly distributed, then X is logarithmic in the general sense, and is of the form  $k/x$ . Note that the fact that X is distributed as  $k/x$  follows directly from Proposition II. Moreover, X is logarithmic according to the General Law since its logarithm is uniform and spans an integral length, implying that mantissa is uniform as well, and thus the General Law holds. That first digits of X are distributed as  $\text{LOG}(1+1/d)$  can be simply deduced from Proposition II implying that the density form of X is  $k/x$ , and utilizing what was observed in the earlier chapter about  $k/x$  distribution being logarithmic in the first-order sense whenever exponent difference is an integer.

It is important to note that the integral restriction on exponent difference of Corollary I and Corollary II becomes redundant and can be ignored if exponent difference is quite large. How large? Well, that depends on the desired precision. Certainly, a fairly large value of (non-integral) exponent difference of, say, over 25, yields digit distribution that is quite close to the logarithmic for all practical matters. The reason for this waiver is that for a very large interval of related log, say a uniform log density over  $[5.0, 37.2]$ , that extra piece of data stemming from the fractional part of log  $[37.0, 37.2]$  over and above the integral interval  $[5.0, 37.0]$  becomes exceedingly tiny as a fraction of overall data. Since  $[5.0, 37.0]$  generates perfectly logarithmic data, contaminating it with a tiny non-logarithmic extra data set of log over  $[37.0, 37.2]$  doesn't disrupt logarithmic behavior by much. Yet, that extra small piece ruins the chance for a perfect logarithmic behavior.

**Proposition III:** If data or distribution is defined between any two adjacent/consecutive integral powers of ten, such as 1&10, 10&100, 100&1000, and so forth, then the only distribution consistent with the general form of Benford's Law is of the form  $k/x$ . Since no other distribution could satisfy the general logarithmic condition, the very definition of being logarithmic in the general sense over such an interval is being  $k/x$ . Uniqueness of  $k/x$  at least over such an interval grants it a prominent role in the field of Benford's Law.

Let us prove this assertion. First we shall assume logarithmic behavior and then show that it is uniquely distributed as in  $k/x$ . If data spans two adjacent IPOT then this implies that its related log is stuck narrowly between two consecutive integers. For example, if data falls exclusively on  $(10, 100)$ , then its related log spans  $(1, 2)$ . Whenever log spans two adjacent integers on the interval  $(N, N+1)$ , both log and mantissa correspond exactly, except for that irrelevant characteristic whole number. Since logarithmic behavior requires uniformity of mantissa, it follows that related log itself here is also required to be uniform in order to obtain general logarithmic behavior. Put another way, within such narrow unity range of log, the only way mantissa could end up being uniform is for the log itself to be so. Uniformity of related log density in turn implies that data is distributed in the form of  $k/x$  as per Proposition II. Same argument (in reverse) is made if  $k/x$  is assumed, since then by Proposition I its related log is uniform, implying that mantissa is uniform as well and thus that data is logarithmic according to the general law.

**Proposition IV: If data or distribution is defined between any two values whose exponent difference is unity, that is whenever interval is on  $[10^S, 10^{S+1}]$ ,  $S$  being any real number, then the only distribution consistent with the general form of Benford's Law is of the form  $k/x$ .**

The same reasoning as for Proposition III applies here, since uniformity of mantissa implies uniformity of related log over  $[S, S+1]$  and vice versa. As an example, any distribution defined on  $(3.47, 34.7)$  namely on  $(10^{0.54033}, 10^{1.54033})$  is logarithmic if and only if it's distributed as  $k/x$ .

On the other hand, for an interval standing between two **non-adjacent** integral powers of ten such as  $(10, 100000)$  for example, no distribution can claim uniqueness. There are in theory infinitely many possible logarithmic distributions on such 'wider interval', and  $k/x$  is but one such manifestation of a density behaving logarithmically there. It is conceivable even to have low digits losing a bit on the leftmost sub-interval  $(10, 100)$ ; a near logarithmic behavior around the center on  $(100, 1000)$  and  $(1000, 10000)$  say; and some very strong leadership of low digits — even stronger than the Benford condition — on the rightmost sub-interval of  $(10000, 100000)$ , to compensate for their previous loss on the left; all done and carefully calibrated in such a way as to leave aggregate leading digits distribution perfectly Benford!

Let us end this chapter by noting two important aspects of  $k/x$  distribution. The first aspect is that concentration is constant between any two adjacent IPOT points. For example, for  $k/x$  defined over  $(1, 1000)$ , areas on  $(1, 10)$ ,  $(10, 100)$ , and  $(100, 1000)$  are all equal. If areas represent approximately the number of discrete values falling within them, as in histogram construction of a finite data set roughly along  $k/x$  lines, then it could be said that there are as many values in the interval  $(100, 1000)$  [curve is very low but range is long] as there are in  $(1, 10)$  [curve is very high but range is short]. This fact resonates well considering that related log density of  $k/x$  which is uniform throughout, never diminishes, and is constant across all integral points. To demonstrate this simple fact, we definite-integrate the generic  $k/x$  distribution between any of its IPOT adjacent values:

$$\int k/x \, dx \text{ over } [10^{\text{INTEGER}}, 10^{\text{INTEGER} + 1}] = k[\ln(10^{\text{INTEGER} + 1}) - \ln(10^{\text{INTEGER}})] = k[(\text{INTEGER} + 1) \cdot \ln 10 - (\text{INTEGER}) \cdot \ln(10)] = k \cdot \ln 10$$

which is a constant and independent of the value of INTEGER. In fact, this result is by far more general and does not depend on the value INTEGER being an integer, or even that unity exponent difference exists (the exact value of 1), as can be clearly deduced from the proof above. The distribution  $k/x$  has a constant area for all of its sub-intervals standing between any fixed constant exponent difference  $F$ . In other words,  $\int k/x \, dx \text{ over } [10^G, 10^{G+F}]$  is a constant  $k \cdot \ln(10) \cdot F$ , for all values of  $G$  without requiring  $G$  to be an integer or for  $F$  to be 1.

Secondly, let us demonstrate now directly what was claimed in an earlier chapter, namely that whenever the distribution is defined over a long interval of say  $[10^1, 10^{15}]$ ,  **$k/x$  can be considered ‘continuously logarithmic’**. Since its related log is uniform throughout the entire range of  $[1, 15]$  on the log-axis as per Proposition I,  $k/x$  over say  $[10^1, 10^2]$  is also perfectly logarithmic since its related log on  $[1, 2]$  is uniform just the same! But so is any internal sub-interval between any two adjacent IPOT points such as  $[10, 100]$  or  $[1000, 10000]$ , since log is uniform anywhere within  $[1, 15]$ . In fact between any two points with an integral exponent difference such as  $[43, 43000]$ ,  $[2, 2000]$ , or  $[10, 1000000]$  for example, related log is uniform just as well and thus any such sub-section is perfectly logarithmic as per Corollary I. Yet, over sub-sections with non-integral exponent difference such as  $[3.0, 3.2]$  or  $[9.0, 13.8]$   $k/x$  is not logarithmic.

## RELATED LOG CONJECTURE

---

---

Must log density itself be uniform in order to bequeath uniformity to mantissa and consequently Benford's Law to the data? Not necessarily! Surprisingly, log may curve upon itself and uniformity of mantissa can still be found. Little reflection is needed to realize that uniformity of related log over an integral log-axis interval (such as in the  $k/x$  case) is not the only circumstance yielding uniformity of mantissa. In situations where density curve of log of data ('related log') is continuous, has wide enough range on the log-axis (roughly over three), and where it monotonically rises on the left from the log-axis itself until it reaches a certain plateau or zenith, followed by a monotonic fall on the right all the way back to the log-axis, it is conjectured that uniformity of mantissa may also be achieved, at least approximately, especially when it's nearly symmetric and having gradual (not too steep) rise and fall. It all boils down to the question of whether or not the fractional part of a variable can be uniformly distributed without parental variable itself being uniform.

In a case such as an upside-down U-shaped-like or V-shaped-like log density, it monotonically increases up to a central point (meaning that high digits are benefiting somewhat as compared with the logarithmic condition) and then steadily falls off from there (meaning that now low digits get an advantage over and above the logarithmic condition), and it is easy to envision overall mantissa ending up uniform, assuming that the curve traverses plenty of log-axis distance. It is plausible to argue that whatever low digits lose (in relation to the logarithmic) on the left of the highest point (rising) is exactly what they gain to the right of it (falling), regardless of the location of the center. On the other hand, for a log curve that is too narrowly focused on the log-axis having a small range, there is no such meaningful trade-off. For example, if related log curve is bordered by 7.904 and 8.000, and hence generating mantissa on (0.904, 1.000) exclusively, the resultant digit distribution of the data itself necessarily consists only of digits 8 and 9, regardless of the shape of its log curve hanging above.

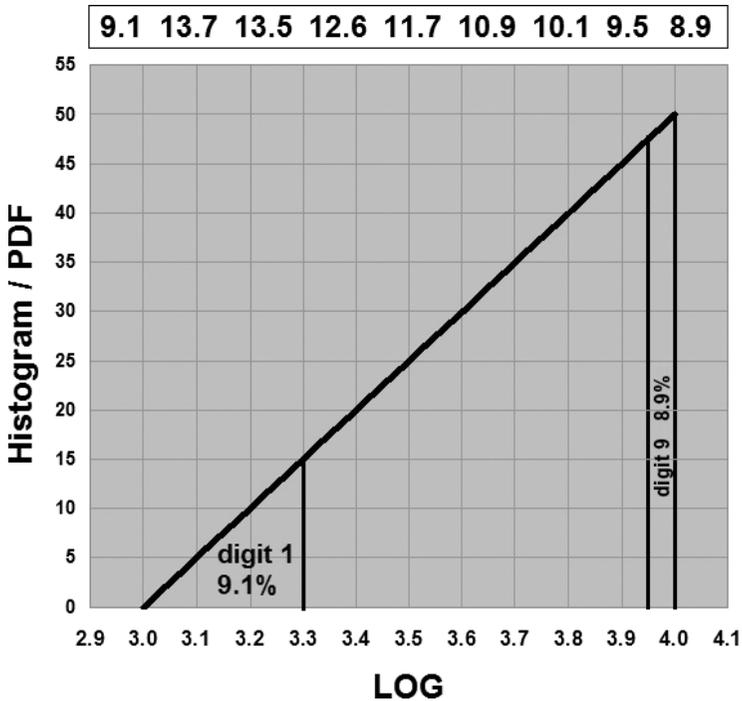


Figure 4.21 Rising Log Yields Non-Uniform Mantissa and Near Digital Equality

Figure 4.21 demonstrates what happens to digit distribution when log is only rising. Since here log is defined on (3, 4) exclusively, log and mantissa are one and the same concept, distributed equally except for that extra 3 in front of the fractional part. A near digital equality exists here. The area for digit 9 which corresponds to the narrow mantissa sub-interval [0.954, 1.000) with taller density approximately equals area for digit 1 which corresponds to the wider mantissa sub-interval [0, 0.301] with shorter density.

Figure 4.22 demonstrates what happens to digit distribution when log is only falling. Here again log and mantissa are one and same concept, and are distributed equally except for that integer 3. An extreme digital skewness exists here in favor of low digits. Area for digit 9 which corresponds to the very narrow mantissa [0.954, 1.000) with very low density, is less than a quarter of 1%, while area for digit 1 which corresponds to the wider mantissa [0, 0.301] with much higher density is more than half of the entire data set. Surely if one shifts the focus of the range from, say, (3, 4) to (2.954, 3.954) in Figs. 4.21 and 4.22, two somewhat

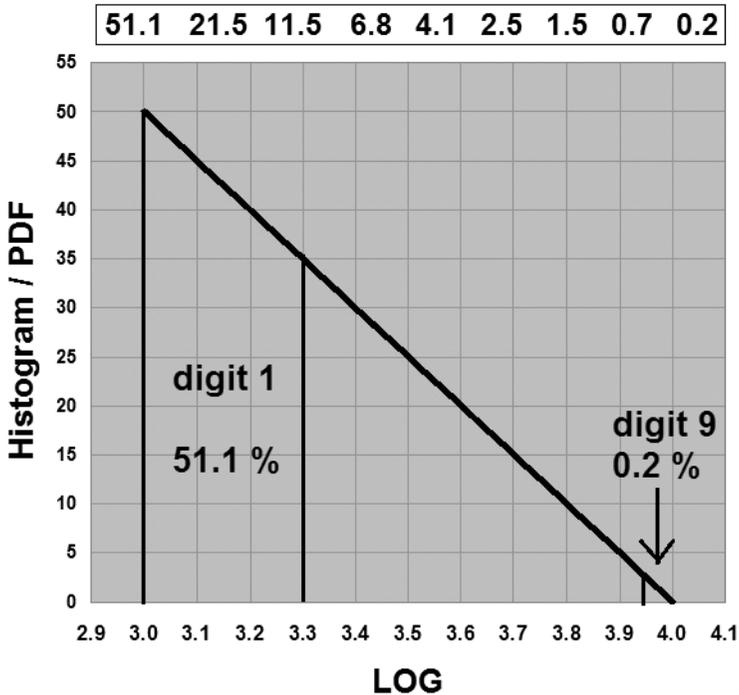
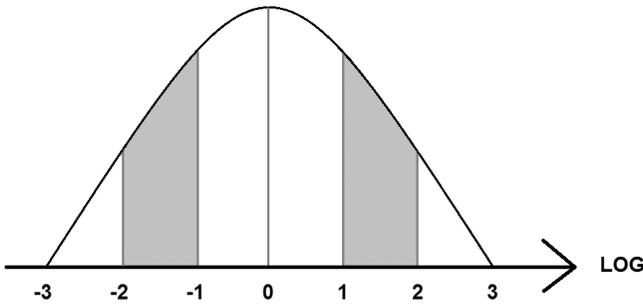


Figure 4.22 Falling LogYields Non-Uniform Mantissa & Extreme Digital Skewness

different configurations of rise and fall emerge, but this doesn't alter the overall conclusion about the resultant relative proportions of all digits 1 to 9, especially when a multiple of such sub-intervals (all falling or all rising) are considered in the aggregate.

Figure 4.23 facilitates the visualization behind related log conjecture. In this graph, log is defined on  $(-3, +3)$ ; it is quite symmetric; it rises and falls gradually; and it spans a wide range of six units on the log-axis. Overall mantissa here is calculated by averaging out (weighted) mantissae from the six different sub-intervals standing between integer points. Let us focus on the two gray areas of  $(-2, -1)$  and  $(1, 2)$ . It is quite plausible to argue that combining mantissae generated by these two separate regions — superimposed one upon another — yields uniformity of resultant mantissa [this is so in light of the fact that no data weighing is necessary where both regions are of equal data proportion]. The same comparisons and conclusions can also be made between the regions  $(-3, -2)$  and  $(2, 3)$  as well as between the regions  $(-1, 0)$  and  $(0, 1)$ . Since such moderations and balancing exist



**Figure 4.23** Non-Uniform Log Curve That Yields Uniformity of Mantissa

separately within each of the three pairs of regions here, overall mantissa [for all six regions combined] is seen as being uniform in the aggregate.

Figure 4.24 shows the stripes on the log curve where digit 1 leads. The curve is actually the Normal distribution with mean 0 and standard deviation 1.65, which was chosen for the purpose of easy illustration. The black stripes are simply locations on the log-axis where mantissa is on  $[0, 0.301]$ , namely on intervals such as  $[N.000, N.301]$ , with  $N$  being any integer. On the left of the origin (negative side) the stripes appear a bit less than a third of each area over  $[N, N+1]$ , while on the right of the origin (positive side) the stripes appear a bit more than a third of each area over  $[N, N+1]$ . Some grand trade-off and cancellations between the left and the right sides are expected. The fact that the overall proportion of these black stripes seems to be about a third of the entire area over  $[-5, +5]$  confirms the general intuition of related log conjecture. The construction of these black stripes was based on the left third segment of each unity length on the log-axis. If instead, white stripes are constructed on each  $[N.954, N.999]$  log-axis segment, corresponding to digit 9 leading, then they all would appear approximately as occupying 4.6% of the entire log area in the aggregate.

The chart in Fig. 4.25 shows a hypothetical density curve of related log. Density curve starts from the very bottom on the log-axis, touching it, and ends all the way down there as well. It also spans plenty of log-axis range (over nine units), thus its related data can be said to have a very large order of magnitude. On the extreme left near  $-2$  or  $-1$ , high digits in a rare show of force are winning slightly, then digital equality is achieved at a single point perhaps, and it is approximated here to be so on a wide interval for better visibility. Further on towards the center, low digits win slightly, followed by logarithmic behavior around the center.

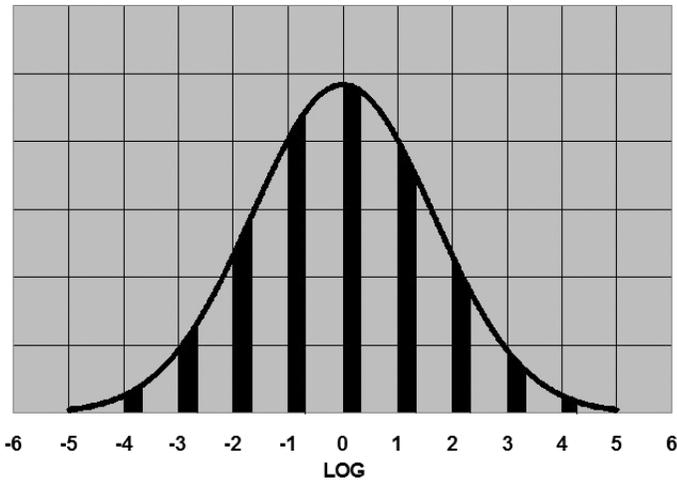


Figure 4.24 Locations on Non-Uniform Log Density Where Digit 1 Leads

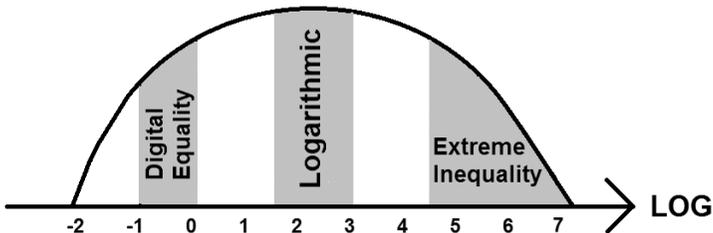


Figure 4.25 Log of Data and Its Digital Status Throughout

A bit farther to the right, low digits win even more than that, and so forth. Gradually we end up on the extreme right where low digits almost completely dominate all other digits, much more so than their normal logarithmic advantage. Intuitively all this suggests some grand trade-off between the left region and the right region, leaving the logarithmic center as a good representative of the entire range. The chart in Fig. 4.25 is not some invented fantasy concocted in the excited mind of the bold and imaginative statistician, but rather an extremely typical log curve found in almost all random data sets, including physical data handmade by Mother Nature herself who strongly favors such round logarithmic curves.

## TESTING RELATED LOG CONJECTURE VIA SIMULATIONS

---

Let us examine related log conjecture via computer simulations. Three obvious examples for symmetrical log distributions come to mind: (1) The Normal distribution, (2) the triangular, and (3) the semi-circular.

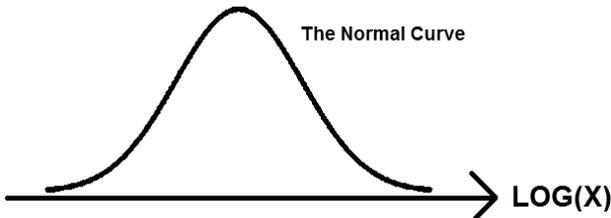


Figure 4.26 The Normal as Related Log of Data

Figure 4.26 depicts the arrangement of the Normal as the density curve of log of data  $X$ . Simulations show that if related log of data is normally distributed, then data itself, namely  $10^{\text{Normal}}$ , is nearly logarithmic as long as the standard deviation (s.d.) of the normal is approximately 0.4 or larger, regardless of what value the mean takes. Increasing s.d. to approximately 0.7 increases agreement with the logarithmic accordingly; and the value of 1 for the s.d. is quite sufficient for a near perfect agreement. The reason for the requirement of a large value for the s.d. is that it guarantees wide enough spread over log-axis, while too small a value leaves the distribution too narrowly focused on some very small interval. The value of the mean does not influence at all the length of the range or the shape of the Normal curve, thus digital behavior of related data is almost indifferent to it for high s.d. values. By Chebyshev's theorem, 8/9 of any distribution is within 3 s.d. of the mean. Applying this to the above mentioned s.d. value of 0.4, we then have roughly 89% of the distribution over a range of at least 2.4 units  $[0.4 \cdot (3+3)]$  on the log-axis. For the value of 0.7 of the s.d. we have 89% over 4.2 units

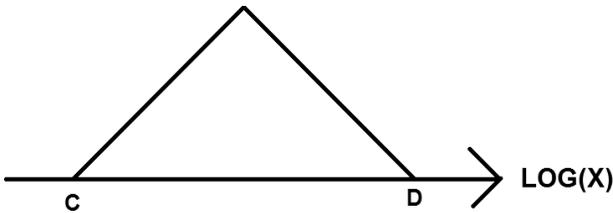


Figure 4.27 The Symmetrical Triangular as Related Log of Data

[ $0.7 \cdot (3+3)$ ] on the log-axis, which is more than sufficient. This particular arrangement outlined above where the Normal serves as the log of data is actually something quite familiar in statistics, namely the Lognormal distribution, except for the different base used there. The Lognormal is defined as  $Y = e^{\text{Normal}}$  while in our simulations  $X = 10^{\text{Normal}}$  was constructed, yet, as is known in mathematical statistics, the particular choice of the base is immaterial and  $10^{\text{Normal}}$  is Lognormally distributed just the same.

Figure 4.27 depicts the arrangement of the symmetrical triangular as the density curve of log of data  $X$ . Simulations of these symmetrical triangular distributions use the following programming commands to obtain a random  $\text{LOG}(X)$  value by way of basic geometric principles:

```
IF RAND < 0.5 THEN: C + SQRT((RAND/2)*((D - C)2));
ELSE: D - SQRT(((1-RAND)/2)*((D - C)2))
```

Here  $D$  and  $C$  are the upper and lower bounds respectively;  $\text{RAND}$  is a simulated number from the Uniform on  $(0, 1)$  representing a particular value for the cumulative distribution function; and  $\text{SQRT}$  is the square root function. The maximum height at the center, namely at  $(D + C)/2$  is uniquely determined by the values of  $C$  and  $D$ , since total area must add up to unity, therefore references to the height are omitted. These simulations yield strong logarithmic results for  $10^{\text{Triangular}}$  whenever  $(D - C) > 2$ . In other words, for those triangular log distributions whose ranges are longer than two units on the log-axis. Longer ranges yield better results and a range of over four or five yields nearly perfect logarithmic behavior. Crucially, the strong logarithmic behavior here for  $(D - C) > 2$  holds almost always regardless of the location of the center; that is, results are almost independent of the value of  $(D + C)/2$ , and depends almost solely on the value of  $(D - C)$ , that is, on the width. For ranges that are narrower than two, fractional translation of location does matter. If the center  $(D + C)/2$  stands exactly on an integer, and if entire range is such that it also spans exactly integral units to the

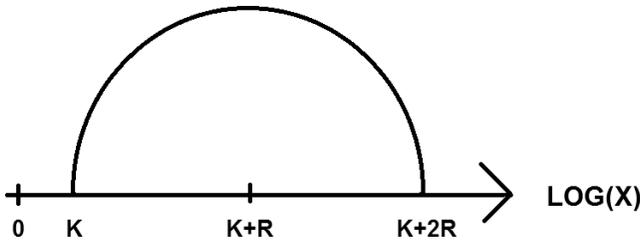


Figure 4.28 The Semi-Circular as Related Log of Data

right of the center and (equal) integral units to the left of it, then it could be shown geometrically that overall mantissa is exactly uniform, that is, that the rise on the left and the fall on the right cancel each other exactly. As an example, the symmetrical triangular  $D = 8$  and  $C = 2$ , center = 5, is perfectly logarithmic.

Lawrence Leemis claims in communicating with the author that for any related log density distributed as  $\text{Triangular}(C, \text{max}, D)$ , (asymmetrical as well); max being the log-axis value yielding the highest point on the triangular (the mode), leading-digits distribution of  $10^{\text{Triangular}}$  is exactly logarithmic if  $C$ , max, and  $D$  are all integers. This exact result is claimed regardless of the value of the entire range ( $D - C$ ), including the lowest possible length of a mere two units, namely  $\text{Triangular}(C, C+1, C+2)$ , as in  $\text{Triangular}(3, 4, 5)$ . The author could not independently verify his claim, except for symmetrical triangles positioned exactly onto the integers.

Figure 4.28 depicts the arrangement of the semi-circular-like distribution as the density curve of log of data  $X$ , with the diameter laying on the log-axis. Computer simulations of the semi-circular-like distribution are quite difficult to perform due to the form of the density function here; hence they are abandoned in favor of direct calculations. The actual shape of the distribution here should not be truly circular, since area under the curve must sum to 1, whereas area of the semi-circle is  $\frac{1}{2}\pi R^2$  and we seek total flexibility in choosing any value for the radius  $R$ , not only the unique solution to  $\frac{1}{2}\pi R^2 = 1$ . There is a need to introduce the adjusting factor  $(\frac{1}{2}\pi R^2)^{-1}$  leading to  $(\frac{1}{2}\pi R^2)^{-1}(R^2 - (\log(X) - (K+R))^2)^{(1/2)}$  as the expression for the density function of  $\log(X)$ . Hence the adjusting factor bends and contorts the circular area downward for large  $R$  values.

Direct calculations of the semi-circular-like distributions as log of data shows satisfactory logarithmic behavior for  $10^{\text{Semi-Circular-Like}}$  whenever  $R$  is roughly larger than 1.1, namely a total range of over 2.2 on the log-axis. A near perfect agreement with the logarithmic is observed for large enough value of  $R$ , say over three. Note that for large enough value of  $R$ , almost the same digital result is obtained whenever center of distribution is displaced to the left or to the right by

any amount; hence the idea presented here is quite general and independent of the specific location on the log-axis.

Taking stock now, we note that in general, a range of roughly 2.5 was quite sufficient for the normal, triangular and semi-circular-like related log distributions to bequeath near-logarithmic behavior to the data. Moreover, location on the log-axis did not seem to matter, so that shifting the whole curve to the left or to the right by any fractional amount was harmless to logarithmic behavior for the most part. Certainly, any left or right translation of related log curves by an integral value does not change digit distributions of the data in any way, since mantissa is totally unaffected by such whole-number shifts. Therefore, it is conjectured that this 2.5 value is quite general and should also be sufficient when applied to other symmetrical log densities. A very large range of, say, four or five log-axis units is conjectured to yield near-perfect logarithmic behavior in general for (almost) all distributions — so long as they start and end on the log-axis itself and curve themselves gradually.

One should not lose sight of the distinction that must be made between the entire span on the log-axis as recorded in all three simulations above, akin to the Order of Magnitude (OOM), and the more general and applicable definition of Order of Magnitude of Variability (OMV) defined earlier as  $\text{LOG}(90\text{th percentile}) - \text{LOG}(10\text{th percentile})$ . Eliminating 10% on either side in all of the above simulations would yield a somewhat lower minimum OMV value necessary for logarithmic behavior, namely that approximately only 2.2 OMV is required in all the above simulations in order to obtain a strong logarithmic behavior in the data.

It should be noted that  $e^{\text{Related Log}}$  demands much larger range on the (common/decimal) log-axis in order to bequeath logarithmic behavior to data than what  $10^{\text{Related Log}}$  usually demands. For example, simulations for related log in the form of the symmetrical triangular distributions require a range of only about two units on the log-axis for a reasonable logarithmic behavior of  $10^{\text{Symmetrical Triangular}}$ , while for  $e^{\text{Symmetrical Triangular}}$  they require roughly four units to behave about logarithmically (double that amount of two). The difference in the log-axis length requirement for bases 10 and  $e$  can be clearly explained in terms of resultant overall variability (Order Of Magnitude). The value of  $e$  is about 2.718282, thus  $10 > e$ , and variability and range of possible values are by far larger for the decimal base than for the natural base. For example, for the exponent set  $\{0.5, 1.0, 3.0\}$ ,  $10^{\{0.5, 1.0, 3.0\}}$  is spread over the wider range of  $\{3.2, 10, 1000\}$ , while  $e^{\{0.5, 1.0, 3.0\}}$  is spread over the shorter range of  $\{1.6, 2.7, 20.1\}$ . The default term ‘related log’ used in this book always refers to the decimal log base 10, not the natural base  $e$ .

## THE LOGNORMAL CONJECTURE OF HILL'S SUPER DISTRIBUTION

---

Let us examine related log of one realization of Hill's super distribution in the context of Related Log Conjecture. A crude simulation of a mini (and very finite) Hill's super distribution was performed. It employed just six different distributions and a throw of a regular unbiased six-side dice 20,000 times to decide which to draw from. Admittedly, this is a very crude attempt in investigating further the statistical structure within Hill's scheme; yet since some other trials in the same spirit gave similar results, this line of attack shall be pursued. Care was taken to avoid distributions with negative values which are forbidden in his model. Below is the list of these six distributions. [The notation  $U(0, b)$  signifies the Uniform on  $(0, b)$ .]

$$\text{DIST}_1 = U_1(0, 2) + U_2(0, 0.3) + U_3(0, 3)$$

$$\text{DIST}_2 = U_1(0, 3.8) + | -3.8 * \ln(1 - U_2(0, 1)) - 7 |$$

$$\text{DIST}_3 = | \text{Normal}(3, 8) |$$

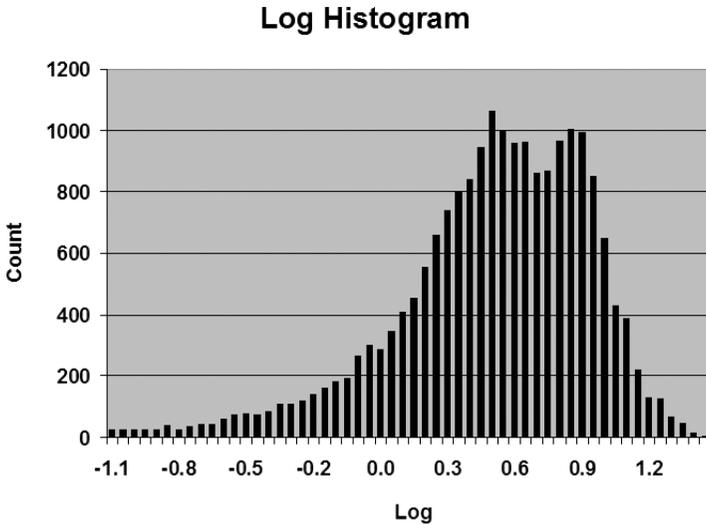
$$\text{DIST}_4 = | U_1(0, 15) - U_2(0, 13) |$$

$$\text{DIST}_5 = | U(0, 1) - \text{Normal}(1, 7) |$$

$$\text{DIST}_6 = | U_1(0, 7) - U_2(0, 5) |$$

First-digits distribution of this simulation did not come out logarithmic of course, rather it came out as  $\{23.7, 16.7, 14.4, 10.5, 8.2, 7.6, 7.6, 6.1, 5.2\}$ , which is almost nicely monotonically decreasing, but having some substantial deviation from the logarithmic. Yet, in spite of the fact that digits here are not close enough to the logarithmic, this result constitutes one quite remarkable feat by Hill and a nice demonstration of the ability of his model to turn around totally non-logarithmic distributions into something fairly close to it by simply mixing merely six of them in a random fashion!

Figure 4.29 is the histogram of the log (base 10) of the above simulated example from Hill's super distribution of 20,000 values. Firstly, it is noted that the entire range on the log-axis (OOM) is a bit narrow, roughly from  $-1.0$  to  $1.3$ , spanning about 2.3 units of length, which perhaps explains why it is not close



**Figure 4.29** Log Histogram of a Mini Hill's Super Distribution Simulation Scheme

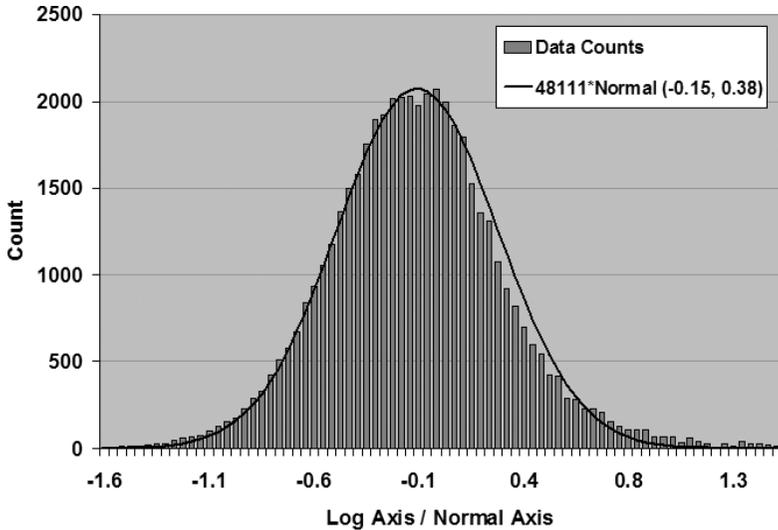
enough to the logarithmic. In terms of OMV the span is calculated approximately (more conservatively) from, say,  $-0.5$  to  $1.1$ , yielding the low value of  $1.6$ . Secondly, the histogram touches the log-axis, both where it is launched on the left and where it ends on the right. Thirdly, it is nicely continuous. Fourthly, the rise and fall are gradual for the most part, not abrupt. Lastly, and most crucially, the histogram is decisively non-uniform! It curves! Moreover, it might even be said to resemble the Normal distribution in some very limited sense [even though it's not symmetric]. Even though this is a truly crude way of simulating Hill's model, the conjecture is made here that his scheme ultimately leads to the Lognormal distribution, or equivalently, that the log of his scheme is Normal. The conjecture states that with a more elaborate simulation scheme encompassing numerous other distribution forms, each having a wide variety of parametrical values, the Normal distribution should appear in the histogram of the log-transformed realized values. (Note: even though the Lognormal involves taking base  $e$  to the Normal distribution while the graph here is of base  $10$ , all this doesn't matter as they are still equivalent in form).

Is it possible that Hill's model leads to the Lognormal? If piling up such incredible variety of unrelated densities can still yield an orderly and predictable digit distribution, namely the logarithmic, could we then dare to suggest that it yields in the limit some orderly and predictable density curve as well?! If so could it be

the Lognormal? Other density options for his related log density, such as the semi-circular-like or the triangular seem highly implausible and rather arbitrary. Hill's related log certainly cannot be the Uniform (miraculously spanning two very-far-apart IPOT points)! Surely related log must be shaped like an upside-down U curve, and if not as in the Normal, then as in what?! Could the (Multiplicative) Central Limit Theorem play a hidden role here perhaps?! The Multiplicative CLT does serve as the basis of one major aspect of the Lognormal. But then what should the location and shape parameters look like in the limit? Why should any single number get the privilege of serving as a parameter here as opposed to another number? Is it possible that parameters get fixated on some two very 'generic' values in the limit as we add more and more distributions, perhaps 10 being the base, 1 being unity, or even 9 for being (base-1)? Could it involve the geometrical Pi  $\pi$ , or Euler's number  $e$ , or some function of the above possibilities?

Seeking a better empirical evidence for this conjecture, absent mathematical proof as yet, an actual collection of real-life data, 34,269 values in all from over 70 different Internet sources, representing a good variety of topics in the spirit of Hill's scheme was performed. It yields a related log curve with a shape that has a definite similarity to that of the Normal (by far more fitting than Fig. 4.29), and at a minimum not too different from that of the contorted semi-circle-like. But it is certainly not flat in the least. More on this real-life empirical simulation is given in Chapter 110.

Reluctantly, the author wishes to extend the Lognormal conjecture to serve as the density for other non-Hill classic logarithmic data sets of single-issue physical quantities, such as amount flow of rivers, population data, time between earthquakes, and so forth. Here, parameters could be particular, aimed at the specific phenomenon or quantity on hand, and hence less mysterious. This extension of the conjecture is not based on speculative notions, but rather on actual examinations of numerous such real-life data sets, all of which approximately confirmed this suggestion! The reluctance to extend the conjecture springs from the fact that Hill's scheme is a mathematical construct that can be further worked on and extended, and perhaps shown to lead to the Lognormal, while the physical manifestations of Benford's Law are as yet without any mathematical framework, more mysterious, and one has no option here but to resort to mere empiricism. What could be proposed though is to somehow demonstrate that a single-issue physical quantity is built upon numerous multiplicative random factors, in which case the process is Lognormal via the Multiplicative Central Limit Theorem. Yet, it is not at



**Figure 4.30** Log Histogram of Distances from Stars Nicely Fitting the Normal

all clear if one could argue that each earthquake, river flow amount, population count, and so forth, is derived from some set of several multiplicative factors.

A good example of a real-life single-issue physical data set where the Lognormal conjecture appears quite compelling is the data set pertaining to a selection of 48,111 stars in our Milky Way Galaxy. The variable in focus here is the distance from our solar system to any given star in units of 1000 light-years. This very limited set from the much larger collection of known stars was made independently of the distance, meaning that distance was randomly obtained. The data is provided courtesy of NASA and it is found on their website <http://heasarc.gsfc.nasa.gov/db-perl/W3Browse>. First-digits distribution is nearly logarithmic but not close to perfection, perhaps due to a slightly narrower range on the log-axis than what is warranted, being roughly 2.3 units in total (OOM). Specifically first digits are {28.3, 15.1, 12.0, 10.5, 9.0, 7.6, 6.5, 5.9, 5.1}, with a low SSD value of 13.8. The chart in Fig. 4.30 shows a remarkable fit to a superimposed Normal having similar parameters as for those of the log of the data. Probability values for the Normal were multiplied by 48,111 so as to depict the same count on a joint histogram. The mean and standard deviation of the log of the star data are  $-0.15$  and  $+0.44$  respectively. Yet, as it turned out, standard deviation of  $0.38$  for the Normal gave a better fit here. All in all, the compatibility of this data set with the Lognormal is quite compelling!

## NON-SYMMETRIC RELATED LOG CURVES

---



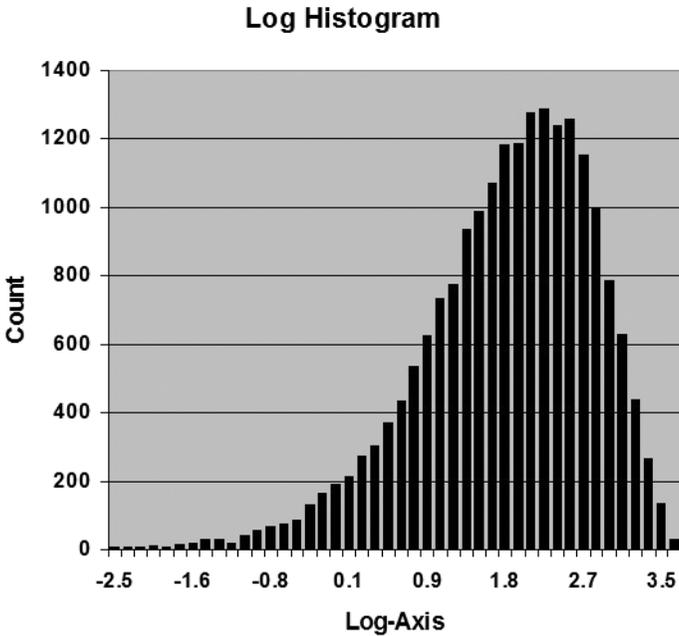
---

A chain of five Uniform distributions was simulated, with the last one being the Uniform on (0, 10000). Schematically the setup of this arrangement can be written as  $U(0, U(0, U(0, U(0, U(0, 10000))))$ ). Resultant leading-digit distribution came out almost perfectly logarithmic. Its related log histogram is not symmetrical, and its tail to the left is longer than the one to the right. In any case, total range is wide enough on the log-axis, at the extremely comfortable five-unit length, compensating perhaps for its asymmetry. It starts and ends on the log-axis itself, gradually without too steep a rise or a fall, and it is nicely continuous. Figure 4.31 depicts the histogram of related log of this particular simulation scheme.

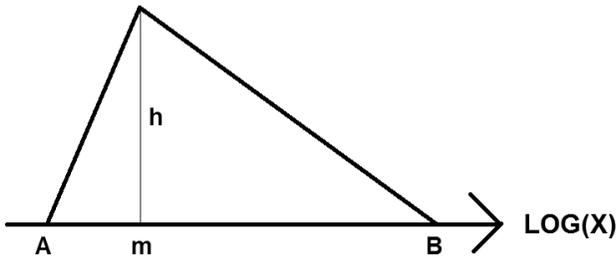
Inspired by such non-symmetrical related log distributions that nonetheless bequeath nearly perfect logarithmic behavior to its data, we are led to explore simulations of an asymmetrical triangular distribution as depicted in Fig. 4.32. This is done using the following programming commands to obtain a random  $\text{LOG}(x)$  value by way of basic geometric principles:

```
IF RAND < (m - A)/(B - A) THEN:
  A+SQRT(RAND*(m - A)*(B - A))
ELSE: B - SQRT((1 - RAND)*(B - m)*(B - A))
```

Here  $B$  and  $A$  are the upper and lower bounds respectively,  $m$  yielding the highest density (the mode),  $h$  the height at  $m$ , and  $\text{RAND}$  being a simulated value from the Uniform(0, 1). These simulations point to logarithmic behavior for  $10^{\text{Asymmetrical Triangular}}$  whenever  $(B - A)$  is roughly larger than 3.5, as long as  $m$  is not too close to either  $A$  or  $B$ . More specifically, whenever  $m$  is not within approximately 0.4 units of distance from either  $A$  or  $B$ , namely whenever the density is not too steep and not too asymmetric. It might be conjectured that here again, just as was seen in the chain of five uniform distributions scheme, the system needs to compensate for its asymmetry by using a slightly wider range (approximately one extra unit on the log-axis).



**Figure 4.31** Log Histogram of a Chain of Five Uniform Distributions

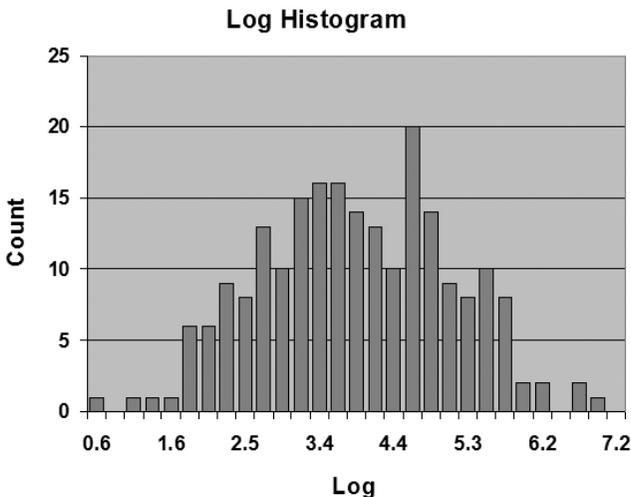


**Figure 4.32** Asymmetrical Triangular Distribution as Related Log of Data

How does it all work out to uniformity of mantissa in spite of its asymmetry? The conceptual explanation is that to the left of point  $m$ , the curve is very steep, so low digits suffer greatly (in comparison with the logarithmic condition) but only very briefly so, while to the right of point  $m$ , density inclines quite gently helping low digits only mildly there, yet doing so for a much longer duration (longer stretch), resulting in that same near perfect trade-off just as was seen for symmetrical log densities.

## WIDE RANGE ON THE LOG-AXIS AND LOGARITHMIC BEHAVIOR

A compelling demonstration of the intimate connection between relatively wide range on the log-axis (i.e. OOM) and logarithmic behavior can be seen in the data pertaining to carbon dioxide emissions by 216 sovereign states and territories in 2008 seen in Chapter 41 on Summation Tests. First digits distribution here is  $\{26.7, 17.1, 10.1, 16.1, 9.2, 6.5, 6.5, 4.1, 3.7\}$ , which is a remarkable feat for such a tiny data set of only 217 data points! SSD value of 62.6 is somewhat moderate. This fantastic near-logarithmic achievement can only be explained by noting the comfortable large range it has on the log-axis, thus driving the point that the spread of related log density (order of magnitude) in the context of Benford's Law does impact leading digits a great deal. Figure 4.33 depicts the log histogram of this small data set, showing the relatively very large range on its log-axis of almost seven units. Such large range on the log-axis easily compensates for lack of smoothness in the shape of the curve of the log density, as well as for the scarcity of values in the data set.



**Figure 4.33** Wide Log Range of Data on Carbon Dioxide Emissions Worldwide

## THE REMARKABLE MALLEABILITY OF RELATED LOG CONJECTURE

---

The remarkable result of related log conjecture, pointing to a resultant uniform mantissa out of almost any log density shape — given the above-stated restrictions — has in its basis some very profound geometrical tendencies and forces, surpassing any knee-jerk reaction of intuition about the possibilities, flexibility, plasticity, and malleability of the mechanizations at work here regarding log-digits interaction. There is no need whatsoever in almost all cases for the tip or the center of the curve to align itself near an integer or away from an integer in order to obtain that near-perfect cancellation of rising and falling mantissa, even though mantissa's cycles oscillate strictly along the integers! The requirement that the curve must contain 'top point' where it is monotonically rising on its left, and monotonically falling on its right (second derivative negative throughout) can be easily relaxed in most cases. Random or haphazard moderate dents that bend the curve are typically harmless to logarithmic behavior for the most part. As for one such example, an artificial (invented) log density with two substantial dents is shown in Fig. 4.34. The construction of the curve was made arbitrarily, without reference to any real-life data. Generously wide range on the log-axis, six units in all, might have been one important factor saving related data from any serious deviation from the logarithmic, and at a minimum, this exceptionally wide range has been a favorable factor in the overall result. First digits are: {30.2, 18.6, 12.8, 9.6, 7.8, 6.4, 5.4, 4.7, 4.3}, a remarkable achievement in spite of that odd curve with two large and sinister-looking dents!

A real-life example of dents in log densities can be found in the data on mass of exoplanets. This provides a physical example of the enormous flexibility and universality of related log conjecture. Figure 4.35 depicts the log histogram of exoplanet mass, showing two significant dents. Remarkably, even though the data contains only 800 points, and in spite of these two large dents, the moderately generous range of 3.4 on the log-axis helps first digits distribution overcome the obstacles and get extremely close to the logarithmic. First digits distribution here is {30.0, 18.8, 11.9, 7.4, 8.0, 7.6, 7.4, 4.6, 4.4}!

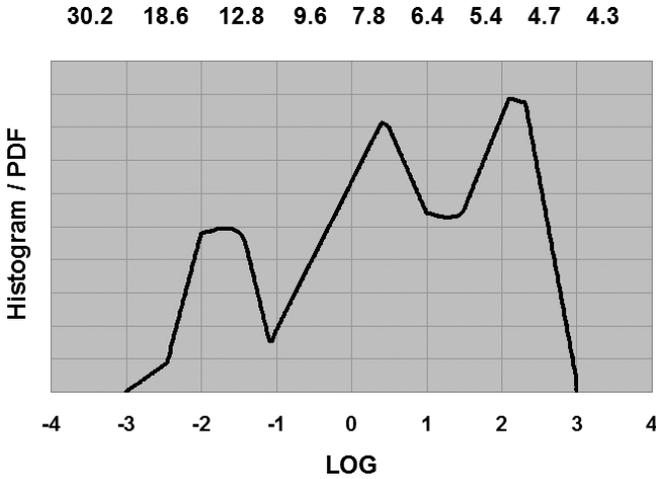


Figure 4.34 The Logarithmic Overcoming Two Severe Dents in Related Log Curve

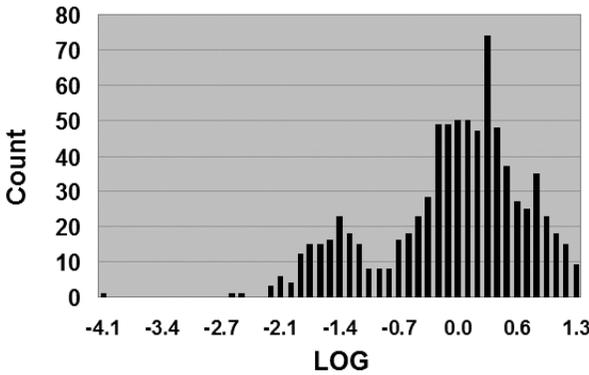


Figure 4.35 Log Histogram of Mass of 800 Exoplanets within the Milky Way

Yet, not having a point dividing monotonically rising left and monotonically falling right could lead to significant deviation from the logarithmic if dents are well-coordinated on the log-axis and arise repeatedly and orderly in relation to the integers on the log-axis. Such artificial or unnatural dents are almost never encountered in real-life data, but rather are man-made in almost all cases, intentionally created to suit some very particular situation or scenario on hand. The chart in Fig. 4.36 depicts one such artificially constructed log density on  $(-5, +5)$ . Resultant first-digits distribution is decisively non-logarithmic and distributed as  $\{23.5, 21.6, 16.3, 11.9, 9.0, 6.6, 4.9, 3.6, 2.6\}$ . Nonetheless it is monotonically

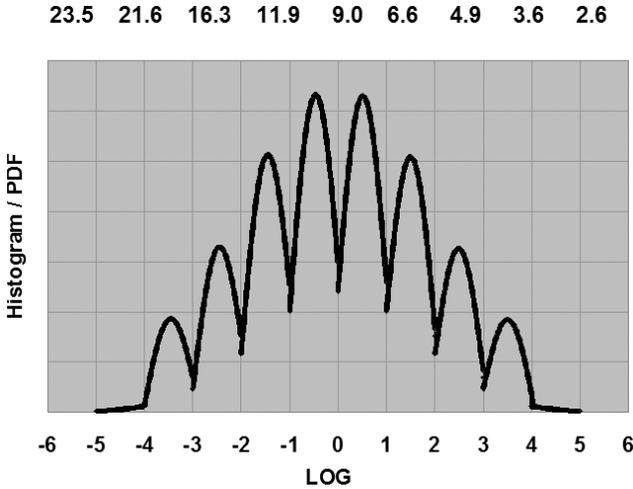


Figure 4.36 Repetitive Dents Aimed at Corresponding Mantissa Locations

decreasing and retains some resemblance to the logarithmic. Here as well, the fact that the curve had relatively very large range ( $\approx 8$  units on the log-axis) must have contributed a great deal in significantly moderating the deviation it made from the logarithmic in spite of these profound and numerous dents.

Yet, if the same wavy shape of curve with those seven dents is constructed on a **narrower** range, say  $(-2, +2)$ , leading digits are now almost fully logarithmic, in spite of all these odd-looking spikes! Calculation of first digits distribution here yields  $\{29.2, 18.0, 13.6, 9.7, 8.0, 5.8, 5.8, 5.3, 4.5\}$ .

And if the same wavy shape of curve with those seven dents is constructed on a **wider** range, say  $(-8, +8)$ , leading digits are again almost fully logarithmic, and first-digits distribution is  $\{30.1, 17.6, 12.5, 9.6, 8.0, 6.7, 5.8, 5.2, 4.7\}$ . Figures 4.37 and 4.38 depict both of these densities on  $(-2, +2)$  and  $(-8, +8)$  respectively. Clearly, digital results of these three different log densities here are not driven by the size of the range, but rather by the degree of coordination of the spikes with integral points on the log-axis. In the first example of  $(-5, +5)$  in Fig. 4.36, there is a perfect coordination and repetition along the integers on the log-axis, where peaks (spikes) are always repeated on the same particular mantissa location, and where valleys are always repeated on another particular mantissa location. Such well-coordinated fluctuations let the spikes and valleys reinforce each other's detrimental effect, culminating in some overall significant deviation from the logarithmic. In contrast, for the two other examples of  $(-2, +2)$  and

(-8, +8) in Figs. 4.37 and 4.38, peaks and valleys are not coordinated at all on the integers, but rather appear to occur randomly and in a haphazardly fashion (integer-wise), leading to an almost total cancellation of effects on overall resultant mantissa, namely leading to an almost perfect trade-off and uniformity of mantissa.

Besides the pitfall of dents, there exists the peril of abrupt launches or terminations. An abrupt launch or termination of the log curve, where there is either a

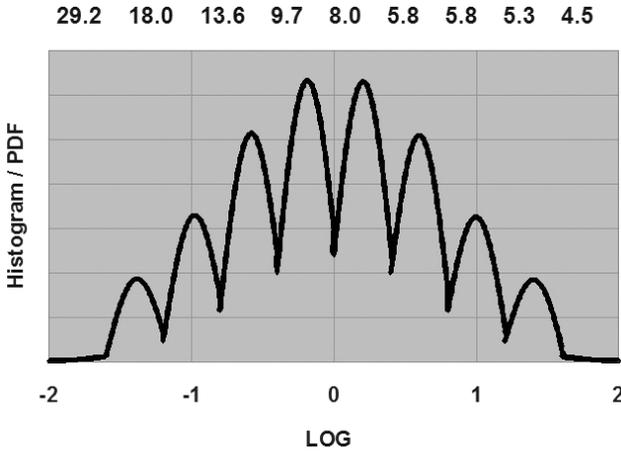


Figure 4.37 Narrow Dents Not Aimed at Any Particular Mantissa Location

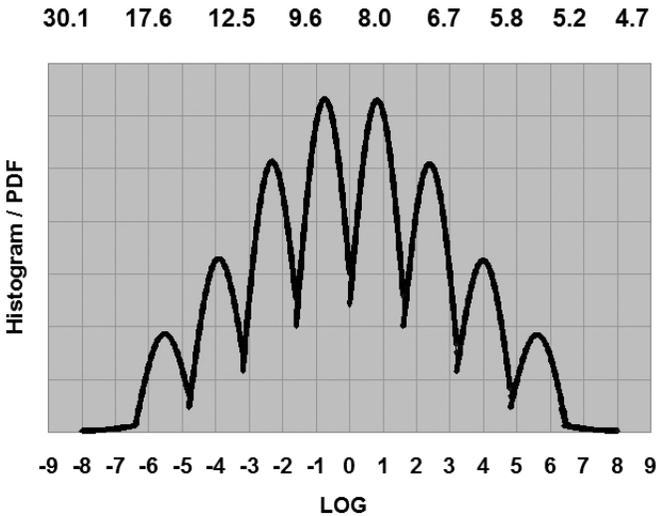
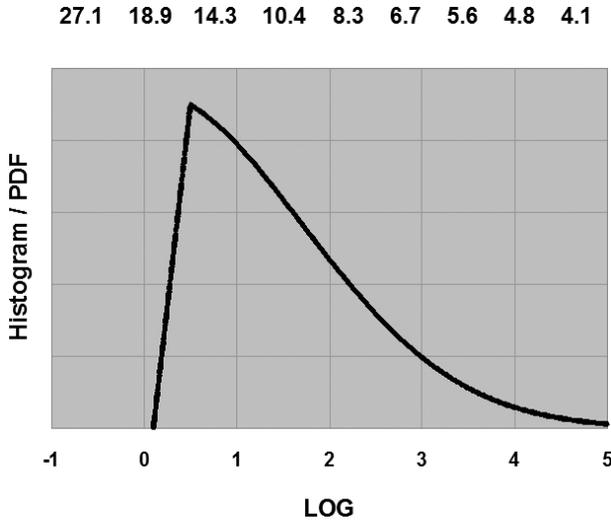


Figure 4.38 Wide Dents Not Aimed at Any Particular Mantissa Location



**Figure 4.39** An Abrupt Log Launching and its Mild Adverse Effect on Digits

steep rise on the left or a sharp fall on the right, is not conducive to logarithmic behavior. Log curve typically needs a gradual rise on the left and a gradual fall on the right in order to bequeath logarithmic behavior to data. An example of what happens to leading digits whenever log suddenly rises is depicted in Fig. 4.39, where it is launched abruptly and then gradually falls. First digits distribution here is  $\{27.1, 18.9, 14.3, 10.4, 8.3, 6.7, 5.6, 4.8, 4.1\}$ ; not close enough to being logarithmic even though log range is sufficiently large. Two extreme cases of abruptness in launching and terminating log densities are presented in Figs. 4.40 and 4.41, where each curve is in fact one side (half) of a Normal with mean 0 and standard deviation 1.7. The curve on  $(0, +5)$  in Fig. 4.40 is being abruptly launched at 0; it's only falling and without any rise as a counterbalance. Therefore its first-digits distribution of  $\{35.1, 18.5, 12.3, 8.9, 7.1, 5.7, 4.8, 4.1, 3.5\}$  is off and skewed in favor of low digits over and above the logarithmic condition. The curve on  $(-5, 0)$  in Fig. 4.41 is abruptly terminated at 0; it's only rising and without any fall as a counterbalance. Therefore its first-digits distribution of  $\{25.3, 16.7, 12.7, 10.3, 8.9, 7.7, 6.8, 6.1, 5.5\}$  is off and less skewed in favor of low digits as compared with the logarithmic condition. It should be noted that the average of these two digit distributions of Figs. 4.40 and 4.41 is almost perfectly logarithmic since as a combination they constitute (for the data itself) Lognormal distribution with parameters that are compatible with logarithmic behavior [their

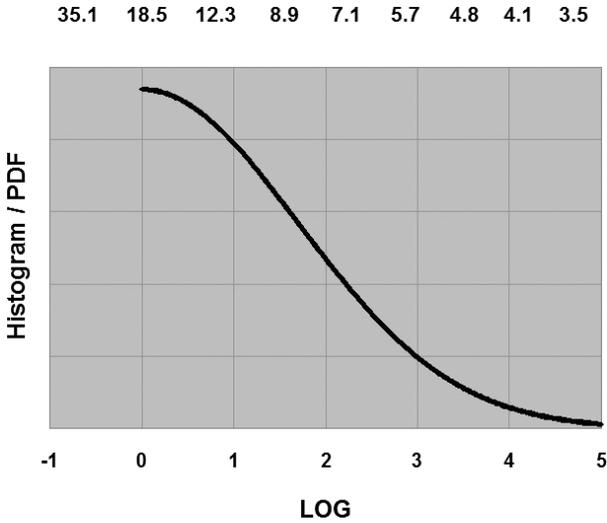


Figure 4.40 A Fall After an Extremely Abrupt Launch and its Digital Effect

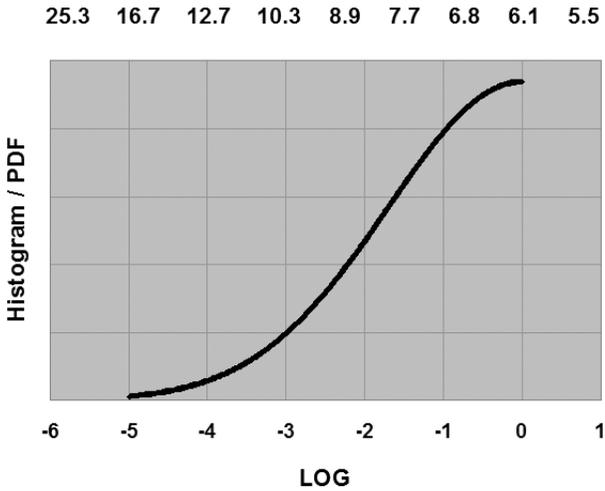
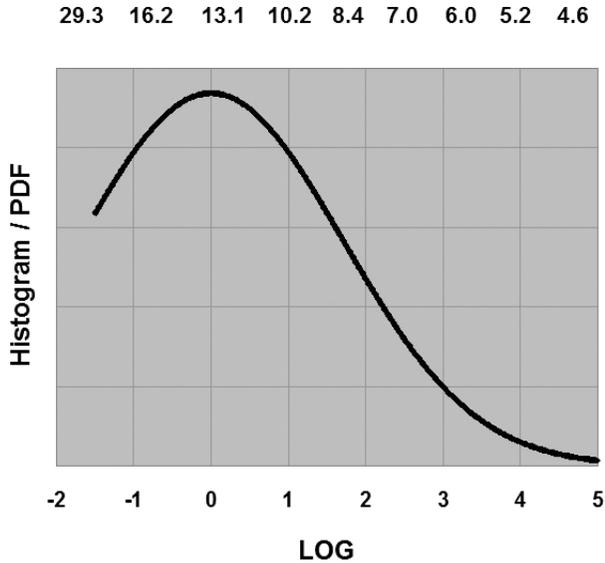


Figure 4.41 A Rise Before an Extremely Abrupt Termination and its Digital Effect

combined log density is Normal with large s.d.]. It must be emphasized though that even extreme cases of abruptness in launching or terminating can easily be moderated or ‘repaired’ by ensuring that there are still two counterpart sections, one section where the curve is rising, and the other section where the curve is falling,



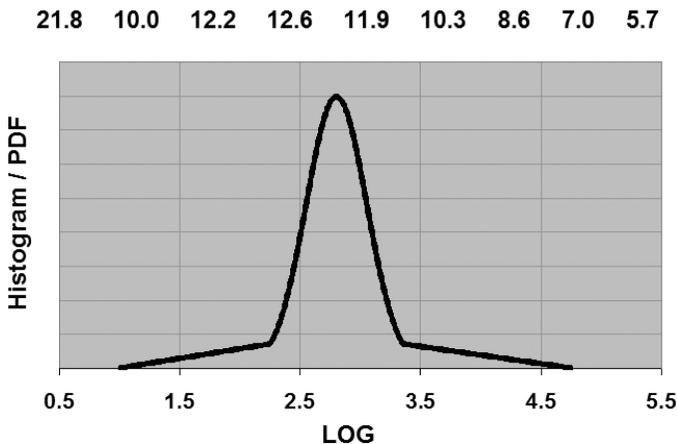
**Figure 4.42** An Abrupt Launch with Counteracting Regions is Nearly Logarithmic

so that at least some moderate trade-off is feasible. The curve in Fig. 4.42 depicts one such case, where the extra section of  $(-1.5, 0)$  on the log-axis added to Fig. 4.40 was enough to yield digit distribution of  $\{29.3, 16.2, 13.1, 10.2, 8.4, 7.0, 6.0, 5.2, 4.6\}$  which is nearly logarithmic. Hence, this addition of a rising curve on  $(-1.5, 0)$  to the falling section on  $(0, 5)$  was for the most part sufficient to balance things out and to lead to a strong logarithmic behavior.

Remarkably, with sufficient range on the log-axis, the damage caused by any abrupt launching or termination is quite limited even for an extreme vertical one, as noted in the above examples where none is anywhere near digital equality or having a digital configuration very different from the logarithmic. This author has labored long and hard in attempting to sabotage as much as possible the logarithmic distribution and to produce worsening digital results with other types of abruptness as well as with those located on other points of the log-axis, all in order to present the reader with a more decisive and convincing example of the logarithmical risk in any abrupt log launching or termination, all to no avail! These failed sabotage attempts demonstrate the plasticity and tenacity of related log conjecture and its ability to generate nearly the logarithmic in most cases (or at least digital configuration not too far from it). Let us recap all the flexible and favorable features we have encountered so far regarding this log-digit interaction: (1) the mild

damage done onto the logarithmic by way of abruptly launching or terminating its log without gradualism, (2) the very permissive attitude towards dents in general, as well as the rarity of any harmful repeated and coordinated dents, (3) the allowance of asymmetrical log shapes, (4) the almost total indifference to where the curve is oriented, i.e. that translations to the right or to the left of the entire log density curve cause minimal or no digital damage whatsoever, and regardless of the shape the log curve, (5) the total indifference to whether or not the entire range on the log-axis spans or does not span an integral value, i.e. that non-integral ranges do not cause any damage whatsoever. These remarkable features of log-digit interaction explain quite well the prevalence, adaptability, and near-universality of the logarithmic property in real-life data sets having sufficiently large order of magnitude. It might be said in conclusion, that **the only real obstacle to logarithmic behavior is insufficient range on the log-axis, namely a small order of magnitude for the data itself**, although in general strong discontinuities and/or highly artificial and unnatural log shapes could become obstacles as well.

**Yet a crucial qualification must be made to the above statement about sufficiency in the length of the range on the log-axis. The estimation and calculation of the length of the range on the log-axis (order of magnitude of data) should not incorporate outliers or even segments of the data that are not part of the main bulk of it, preferably eliminating all together the bottom 10% and the top 10% from the sorted data, and leaving only the central 80% portion of it in the**



**Figure 4.43** Deceptive Wide Range on the Log-Axis When Outliers are Included

**determination of the length.** Figure 4.43 depicts one misleading case where apparent log-axis range spans from 1.00 all the way to 4.75, tempting the novice digital analyst to believe that with such a comfortably large range of 3.75 log-axis units, the data must certainly be logarithmic. Yet this deceptive data set is not logarithmic in the least! Surprisingly, the distribution of the first digits here is actually {21.8, 10.0, 12.2, 12.6, 11.9, 10.3, 8.6, 7.0, 5.7}. To reconcile these two seemingly contradictory facts (i.e. large log range and non-logarithmic behavior), it is necessary to note that the bulk of the data, 84% of it, lies from 2.25 to 3.35, while only 16% resides around it to the left and to the right. This implies that 84% of the data is not logarithmic at all, since its log range spans barely 1.1 units. No matter what type of 16% of data one may offer to mix in, no matter what digital configuration that injected data comes with, overall digital configuration can never be remedied and logarithmic behavior cannot be obtained. Significant deviation from the logarithmic cannot be remedied with a mere 16% of overall data!

The above qualification leads to a better understanding of the mechanizations at play in Fig. 4.39 of abrupt launch. There is no hope of completely remedying the falling region on the right of (0.5, 5) where the bulk of the data lies (89%), with that small rising region on the left of (0.1, 0.5) representing only 11% portion of data. If a milder and less abrupt launch from, say,  $-1.0$  is attempted and succeeds, this would imply that the 'remedying' section on the left of  $(-1.0, 0.5)$  has more substantial portion of overall data, and as such it has the ability to influence overall digit configuration.

The Normal and the Uniform distributions, and their corresponding leading-digit distributions, demonstrate the consistency of the general understanding and theoretical results obtained in this section. Both distributions come with a perfectly symmetrical density curve for the data itself, a condition which yields decisively non-Benford digital distribution as shall be demonstrated in later chapters. Their related log densities easily explain their non-logarithmic behavior in the context of related log conjecture. For Normal distributions on the positive  $x$ -axis, the range on the log-axis is always too narrow. As an example, let us consider computer simulations of a set of 35,000 realized values from the Normal(100, 50). A tiny 2% portion of data, namely 813 negative values are thrown out of the scheme without effecting results much (since log of a negative number does not exist). First digits are {50.8, 5.0, 3.6, 4.7, 5.8, 6.6, 7.2, 7.9, 8.5}. Figure 4.44 depicts the histogram of related log of this Normal distribution. The very narrow log-axis range of roughly 1.2 units is certainly not sufficient to bequeath

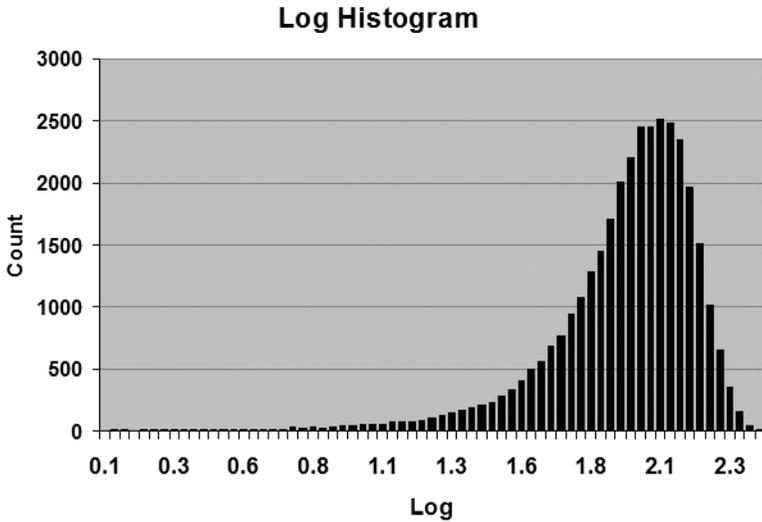


Figure 4.44 Related Log Curve of Computer Simulation of the Normal(100, 50)

logarithmic behavior to the data; and this situation is typical for all Normal distributions with other parameters, so long as they don't crisscross the origin mixing in negative with positive values. The Uniform distribution suffers from different predicaments as its log density is launched abruptly on the left, then rises steadily and (almost) linearly, until an abrupt and quite dramatic termination is encountered somewhere on the right. As an example, let us consider computer simulation of a set of 35,000 realized values from the Uniform(10, 100). Digital equality is expected and digits indeed come out practically uniform at {11.0, 11.1, 11.2, 11.1, 11.1, 11.0, 11.3, 10.9, 11.2}. Figure 4.45 depicts the histogram of related log of this Uniform distribution. The small range of 1.0 unit on the log-axis is certainly not sufficient to bequeath logarithmic behavior to the data. Moreover, the abrupt launch, the dramatic termination, and the steady rise without any offsetting fall, represent other serious impediments to any possible logarithmic behavior. The steady rise throughout the log density of the Uniform and its generic digital equality (subject to the requirement of having IPOT values as endpoints so as to be able to manifest its intrinsic digital-equality feature) corresponds nicely to the generic digital equality associated with particular regions of the rising left in all related log densities as was shown earlier (Figs. 4.21 and 4.25). This consistency in results represents another strong confirmation of the correctness of the log theory developed in this section.

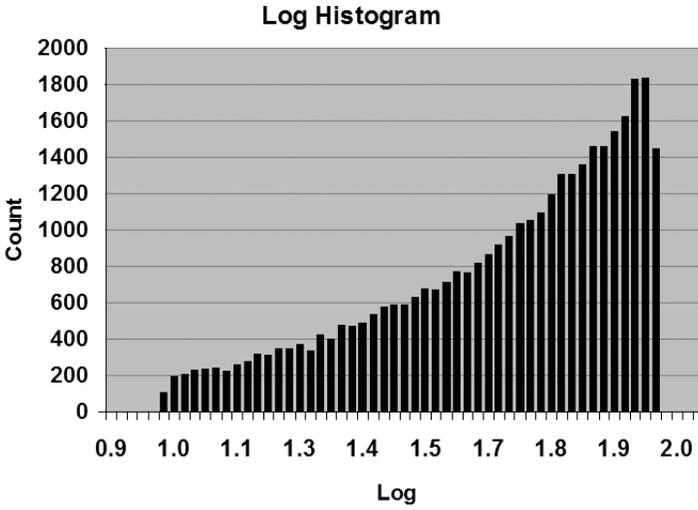
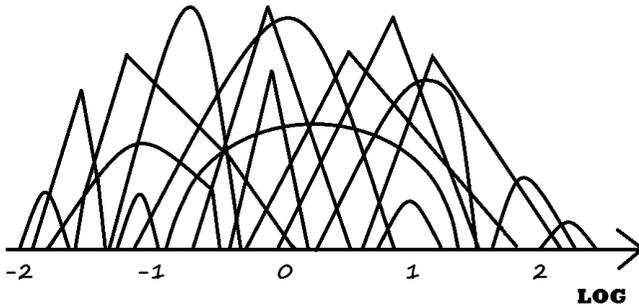


Figure 4.45 Related Log Curve of Computer Simulation of the Uniform(10, 100)

## HILL'S SUPER DISTRIBUTION AND RELATED LOG CONJECTURE

---

Having gained an understanding regarding the remarkable ability of related log in bequeathing uniformity to its mantissa, Hills's super distribution can now be alternatively viewed in terms of a huge collection or aggregation of log densities. Figure 4.46 depicts such a hypothetical collection of curves, not where data histograms of numerous distributions are examined directly, but rather with the point of view of examining their related log curves. Each individual density may not be logarithmic due in particular to an insufficient range on the log-axis, yet in the aggregate the collection of all these short curves gives rise to a singular large density with comfortably huge range. Even if that large curve of the aggregate comes with lots of dents, irregularities, and abruptness (and it shouldn't in the limit), yet as seen earlier, the most crucial factor in logarithmic behavior is the length of the range on the log-axis, and Hill has plenty of that. Discontinuity in resultant curve is discounted in the limit as totally implausible. The assumption here is that all individual curves are confined within some overall finite range, if not of course between  $-2$  to  $+2$  as in Fig. 4.46, then at least within some strict limit. In the application of Hill's scheme to the AGD Interpretation of BL, it can be reasonably assumed that real-life positive data in practical terms comes with some restriction of possibilities on upper and lower bounds. The statement that all real-life (positive) data is confined within the large interval, say,  $[10^{-15}, 10^{+15}]$  seems quite reasonable, and this implies in turn that related log of the aggregation of all real-life positive data (log of AGD over zero) is confined within, say,  $[-15, +15]$ . A cursory look at Fig. 4.46 leads to the very intuitive conjecture that log density of AGD has a gradual rise on the left, a large bulge in the center, and a gradual fall on the right, roughly appearing semi-circular-like or Normal-like. Beyond mere intuition, this claim is supported by some empirical evidence. The scheme mentioned earlier of gathering a large variety of data — 34,269 values in all — from 70 Internet sources focusing on numerous unrelated data sets in the spirit of Hill's model,



**Figure 4.46** Hill's Super Distribution in the Context of Related Log Conjecture

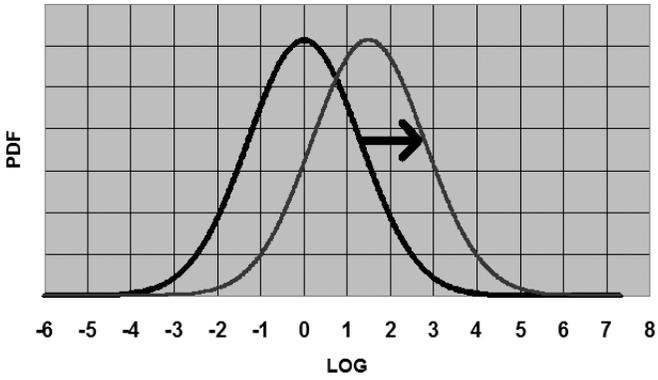
resulted in related log being confined within  $[-4, +10]$ , constituting an enormous yet limited length of 14 units on the log-axis. Figure 6.30 of Chapter 110 regarding this large collection of values confirms the above claims, and corroborates the conjecture about the round shape of related log of AGD. Hill's model may therefore be summarized as simply a scheme of enlarging the span on the log-axis by way of uniting numerous short-spanned densities, that is, enlarging order of magnitude!

## SCALE INVARIANCE PRINCIPLE AND RELATED LOG CONJECTURE

---

An essential feature of related log conjecture, providing for its consistency and stability, is that translations of related log curves to the right or to the left have almost no effect on digits of related data. Translations of log curve by an integral value have no effect on digits of related data whatsoever of course, but even translations by a non-integral value (fraction) in general do not change digital configuration of related data by much given that range on the log-axis is not too narrow. When that range is sufficiently wide and data is decidedly logarithmic, digit distribution of data is nearly totally invariant under any related log translation.

This result is in perfect harmony with the scale invariance principle. In fact, one may state that this **is** indeed the meaning of the principle. A translation on the log-axis of related log curve by the translational value  $+D$  means that each point on the log curve moves to the right by  $D$  log-units. This movement in turn implies that data itself, namely  $10^{\text{Related-Log}}$ , is transformed to  $10^{\text{Related-Log} + D}$ , or equivalently to  $10^D * 10^{\text{Related-Log}}$ , which can be thought of as a scale change by a factor of  $10^D$ . Figure 4.47 depicts a translation to the right on the log-axis of the Normal(0, 1.3) serving as related log of data, by the translational value of 1.5, resulting in translated related log curve of the Normal(1.5, 1.3). Since range on the log-axis was sufficiently large before the translation, it is also so after the translation, and there is almost no movement away from the logarithmic in digital configuration of related data. Translation does not alter shape of the curve, nor does it alter the width of the range below; we are left with the same original length of log-axis units after any translation.



**Figure 4.47** Translations of Related Log Curve Correspond to Scale Changes in Data

## THE NEAR INDESTRUCTIBILITY OF HIGHER ORDER DISTRIBUTIONS

---

The focus throughout all the previous chapters in the discussion about related log conjecture has been on first-digits distribution, of which cyclical length on the log-axis is unity, and where about two to three such cycles are required for logarithmic behavior in the first-order sense. For the second- and higher-orders digit distributions results are much superior in comparison due to their much shorter cycle on the log-axis, requiring by far much shorter ranges for logarithmic behavior in higher-orders sense. For the second order, a full period is cycled each time the first digit changes, hence on the log-axis it is achieved on [C.000, C.301], [C.301, C.477], [C.477, C.602], [C.602, C.699], [C.699, C.778], [C.778, C.845], [C.845, C.903], [C.903, C.954], [C.954, C+1], C being any integer, called The Characteristic [see Fig. 4.19 for illustration]. The fortunate consequence of such short cyclical periods is that it is very rare to find a particular log density curve that manages to distort digit distributions of second and higher orders by much. Given that data obeys BL, second-digits distribution is dependent on first-digits distribution, and vice versa, namely that they have a positive correlation with each other. If a lone number picked randomly from a logarithmic data set is found to have a low first digit, then chances that the second order is also a low digit are slightly higher as compared with the unconditional second order. The unconditional second order is {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}, calculated from the joint probabilities of second order with all first-order digital possibilities. If related log of data is a very short uniform flat line laying exactly on top of one of these nine sub-intervals above of first order such as [C.000, C.301] etc., then second-order distribution varies a bit from its unconditional logarithmic proportions, depending on which first digit rules the territory. If log density is a flat line defined only on [C.000, C.301] where digit 1 exclusively rules the first place, then a skewer condition in favor of low second digits prevails and second-order distribution is {13.8, 12.6, 11.5, 10.7, 10.0, 9.3, 8.7, 8.2, 7.8, 7.4}. If log density is a flat line defined only on [C.954, C+1], where digit 9 exclusively rules

first place, then the least skewed condition of the second order prevails of near digital equality, namely  $\{10.5, 10.4, 10.3, 10.2, 10.0, 9.9, 9.8, 9.7, 9.6, 9.5\}$ . The general view here is that on **all** these nine sub-intervals on the log-axis, second order is always slightly skewed in favor of low digits, but with mild variations in skewness depending on the specific region. Hence any type of aggregations or consolidations of few such regions fused together should all be approximately very near the unconditional second-order digits configurations. This is especially so since offsetting and cancellations between the left side and the right side of the log curve could all be accomplished on a much shorter log-axis range as compared with the case of the first order where relatively larger range of 2.5 units or so was needed. For third and higher orders, the range of a full cycle is even smaller. As an example, for the fifth order, digital equality is almost a certainty, just as our naïve intuition has suggested before Benford and Newcomb, and any deviation from such an equality is a clear indication of outright fraud or some serious error in data, unless the data set is very small, or modeled on some very rare configuration type.

Let us examine second-order distributions for some of the troublesome (first-order-wise) log densities encountered earlier. Not surprisingly, second order has managed to survive quite well all the following calamities: severe dents, coordinated dents, abruptness of all kinds, as well as one very misleading and deceptive log density! When examining these second-order results shown below, they should all be compared with the unconditional second order.

### **Unconditional second order according to Benford's Law:**

**$\{12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5\}$ .**

Figure 4.34, two rather large and menacing dents:

$\{11.6, 11.3, 10.8, 10.6, 9.9, 9.6, 9.4, 9.1, 8.6, 8.8\}$ .

Figure 4.36, wavy dents maliciously coordinated along the integers:

$\{10.4, 10.6, 10.4, 10.5, 10.1, 9.9, 9.9, 9.6, 9.2, 9.3\}$ .

Figure 4.39, abruptly launched on  $(0.1, 5)$ :

$\{11.6, 11.2, 10.6, 10.5, 9.9, 9.6, 9.5, 9.3, 8.8, 9.0\}$ .

Figure 4.40, extreme abruptness in launching; curve is only falling:

$\{12.5, 11.8, 11.0, 10.7, 9.9, 9.5, 9.2, 8.8, 8.3, 8.3\}$ .

Here, very mild deviation from the unconditional is seen, yet it is clearly and consistently skewed in favor of **lower** second digits in comparison with the unconditional second order (just as first order here favors lower first digits over and above

the logarithmic). This is so since there is no trade-off in second-order sense as well, as log curve only falls everywhere having no counterbalancing rise.

Figure 4.41, extreme abruptness in termination; curve is only rising:

{11.3, 11.0, 10.6, 10.5, 9.9, 9.7, 9.6, 9.4, 8.9, 9.1}.

Here mild deviation from the unconditional is seen, yet it is clearly and consistently skewed in favor of **higher** second digits in comparison with the unconditional second order (just as first order here favors higher first digits over and above the logarithmic). This is so since there is no tradeoff in second-order sense as well, because log curve only rises everywhere having no counterbalancing fall.

Figure 4.42, a counterbalanced abrupt launch with two counterparts sections:

{11.8, 11.4, 10.9, 10.7, 9.9, 9.6, 9.4, 9.1, 8.6, 8.7}.

Figure 4.43, a misleading large log-range of (1.00, 4.75), but with bulk of data only on (2.25, 3.35), and the implied low OMV value of  $\approx 1.1$ :

{12.4, 11.6, 10.8, 10.4, 9.7, 9.4, 9.2, 9.0, 8.6, 9.0}.

In conclusion: second order managed to survive quite well all these seven calamities and disruptions in the shape and the range of the log curve and came out almost fully intact.

Figure 4.48 depicts a log density as the Normal with mean 5.8 and standard deviation 0.22. The density of the data itself, namely  $10^{\text{Normal}}$ , is Lognormal with parameters that are not compatible with logarithmic behavior since the shape parameter is too low. Another way of explaining why first digits here are not logarithmic at all is by noting that log-axis range is only about one unit long, namely that order of magnitude is low. Yet second order comes out very nearly logarithmic in spite of it all: {12.6, 11.6, 11.0, 10.0, 9.9, 9.2, 9.1, 9.0, 8.9, 8.8}.

Figure 4.49 depicts a log density as the Uniform on (3.301, 3.602). The density of the data itself, namely  $10^{\text{Uniform}}$ , is  $k/x$  defined on (1999.9, 3999.4) with parameters/range that are not compatible with the logarithmic in the first-order sense, since exponent difference is not an integer. Here digits 2 and 3 totally dominate the first order. Yet, second order is nearly logarithmic. In fact its very slight deviation from the unconditional second order is only due to it being conditional on first digits being 2 and 3. Approximately, second-order distribution here is: {11.7, 11.3, 10.8, 10.4, 10.1, 9.7, 9.4, 9.1, 8.8, 8.6}.

1 <sup>st</sup> :	18.0	7.0	11.6	13.9	14.2	12.0	9.8	7.7	6.0	
2 <sup>nd</sup> :	12.6	11.6	11.0	10.0	9.9	9.2	9.1	9.0	8.9	8.8

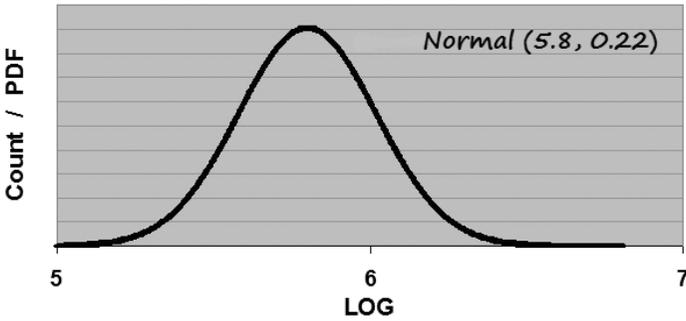


Figure 4.48 Only Second Order is Nearly Logarithmic for Related Data (Lognormal)

1 <sup>st</sup> :	0.0	57.9	42.1	0.0	0.0	0.0	0.0	0.0	0.0	
2 <sup>nd</sup> :	11.7	11.3	10.8	10.4	10.1	9.7	9.4	9.1	8.8	8.6

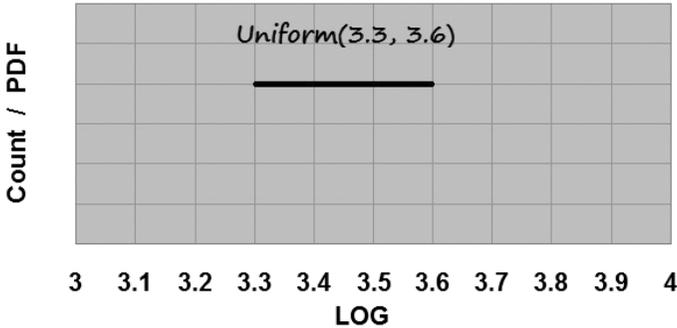


Figure 4.49 Only Second Order is Nearly Logarithmic for Related k/x on (1999, 3999)

Another strong confirmation for the discussion of this chapter is found in Chapter 45, Case Study IX, regarding U.S. County Area data. There, in spite of the fact that data is decisively non-logarithmic in the first-order sense [SSD = 371.8], second order came out quite nearly logarithmic [SSD = 7.7], and third order came out nearly perfectly so [SSD = 0.6]. With each successive higher order having lower cyclical span on the log-axis, we obtain lower SSD! Moreover, even a drastic change of scale from square miles to square kilometers has barely nudged that strong compliance of higher orders with the law [while severely affecting first-order configuration]. Under the Square Kilometer scale, SSD series is: 503.4 for the first, 9.6 for the second, and 2.4 for the third.

## FALLING DENSITY CURVE WITH A TAIL TO THE RIGHT

---

---

In all the simulations of related log densities yielding resultant logarithmic behavior, be it the normal, semi-circular, the triangular, and others, and without a single exception, the histogram of the data itself consistently comes out asymmetrical, with a decisive and sharp fall to the right. In other words, they all come with a clear, long, and prominent tail to the right. In mathematical statistics this property is formally called positive skewness. Yet it must be emphasized that at the launch on the left, the curve is typically rising upward, albeit for a very brief duration, and for a very small portion of overall data.

Interestingly, the same observation about a falling density with a one-sided tail to the right is made in the abstract simulations of Hill's super distribution utilizing six distributions, in the empirical sampling experiment for Hill's scheme from real-life data sets, in all simulated Random Linear Combinations schemes, in all chains of distribution simulations, as well as in all the averaging schemes. Moreover, this feature (falling density) is found in all logarithmic single-issue physical data sets such as time between earthquakes, population count, river flow, pulsar rotation rates, and so forth without a single exception! At the launch on the left there is typically some momentary rise in all of the above-mentioned cases, followed by a much more substantial, long, and consistent fall to the right until the very end of the data on the far right. Standing quite apart from the rabble, proud and pure, the perfectly logarithmic  $k/x$  distribution can claim to have a steady and consistent fall throughout its entire range, as can be read from the algebraic expression of its density function, which is inversely proportional to  $x$ . The distribution  $k/x$  does not suffer from that embarrassing momentary rise at launch.

Is this feature of falling density (in the aggregate) a fundamental condition for all data sets obeying Benford's Law? Is it possible to find logarithmic data sets with a flat histogram, or even with a rising histogram throughout their entire range? In other words, is Benford's Law purely a digital rule, or can it tell us also something about histograms and densities, about the quantitative aspects of typical everyday

data? One must realize that a consistently falling density or histogram implies that curve is high on the left region of the x-axis where values are small, that it is low on the right region where values are big, and that consequently data in the aggregate contains numerous small values, some medium values, and very few truly big values. Such state of affairs gives Benford's Law physical and quantitative dimensions, as opposed to a mere statement about the proportions between the symbolic tools of the language of numbers — digits. In essence, the physical and natural origin of Benford's Law emanates from the fact that typical everyday quantities become more sparse and rarer on higher and higher ranges of the natural numbers, and that small things are much more numerous than large ones.

As a very general argument it might be claimed that if data contains very few high first-order digits such as 7, 8, and 9, and numerous low first-order digits such as 1, 2, and 3, then density is necessarily falling since high digits are always to the right of low digits, as in  $\{1,2,3,4,5,6,7,8,9\}$ . As simplistic as this line of thought may appear, and even in light of a possible counter argument pointing to  $\{6,7,8,9,10,20,30,40,50\}$  as a distinct possibility where high digits are to the left of low digits, the above argument and interpretation of Benford's Law is absolutely correct! [The counter argument neglects to take into account the fact that distance between, say, 3 and 4 is 1, while distance between 30 and 40 is 10.]

Imagine smooth and continuous data that is perfectly logarithmic and limited to an interval between two adjacent IPOT values such as 10 and 100. In this case the digital distribution  $\text{LOG}(1+1/d)$  by definition implies a histogram with a one-sided tail to the right, having numerous numbers in the 10s, and very few numbers in the 90s!

Yet the above example is extremely rare. Everyday data commonly spans multiple IPOT numbers (order of magnitude), and typically comes with very large ranges. Imagine data bounded by two non-adjacent IPOT numbers, such as 1 and 1000. In one scenario, the density falls in all its sub-intervals of (1, 10), (10, 100), and (100, 1000), namely a long tail to the right throughout its entire range. In this case, low digits steadily gain extra leadership at the expense of high digits on each of these sub-intervals standing between adjacent IPOT. The gains over high digits on each of these sub-intervals may vary, depending on how steeply density descends within each sub-interval. Overall digit distribution in this case is made of the aggregate of the mini digit distributions of these three sub-intervals (weighted by portion data within each). In another much more typical scenario, the curve is at times flat or even ascending (typically on the left at launching), and low digits

may lose a few battles, yet they win the war over the entire range in the aggregate, and results come out logarithmic as in  $\text{LOG}(1+1/d)$ . Yet, for low digits to win the war in this last scenario, the histogram of data must come with a substantial fall, much longer and more significant than those momentary rises and flat regions; it needs to fall over the largest portion of overall data.

A good demonstration of the typical manner histograms of (random) logarithmic data sets fall after a brief rise is given by the U.S. 2009 Census Data on population counts of all incorporated cities and towns. This rather large data set of 19,509 such population centers adheres to Benford’s Law very closely. Figure 4.50 depicts only a part of the histogram of this data set, omitting the very long tail portion to the right beyond the 1400 population count (a tail portion which almost steadily falls off from 1400 all the way to 8,391,881). Taking stock, the USA has numerous towns and cities with small population of around 200 or so, few cities with population of 1000 or so, very few large cities with a big population count of around 100,000, and only a handful of mega metropolitans with a population of over a million. The pattern seen in Fig. 4.50 for this population data and its implication about the relative proportions of big and small quantities is a general and ubiquitous attribute across all random data sets that are logarithmic. The significance of the designation “Planet Earth” in Fig. 4.50 will become meaningful in Section 7 of the book, where data is analyzed in the context of other imaginary planets and civilizations having other bases, number systems, traditions, and mathematical abilities, different than the one we have here on Earth.

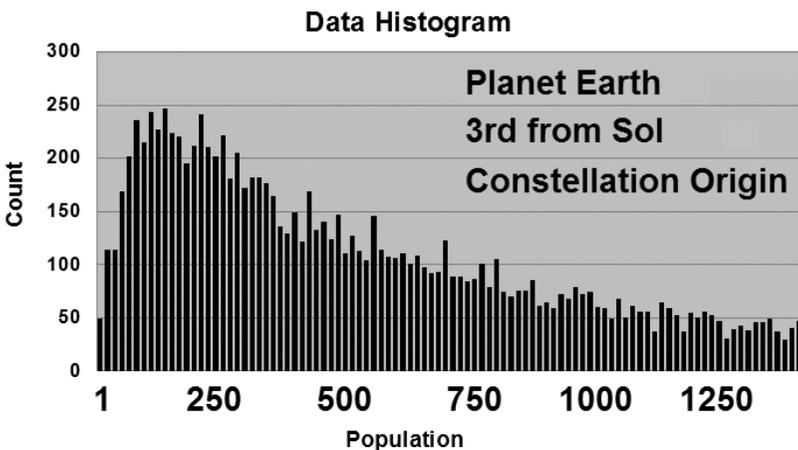


Figure 4.50 Typical Right-Tail-Shaped Histogram of Logarithmic Data — U.S. Population

**It is now sufficiently obvious that all symmetrical distributions, such as the Uniform, the Normal, and others, and regardless of parameters, are always non-logarithmic. By definition they all lack that one-sided asymmetrical tail to the right.** Only a very selective collection of numerous such symmetrical distributions, under particular conditions and arrangements, such as the chain of distributions, Hill's super distribution, and certain others constructs, could end up logarithmic, and only because in the aggregate that overall resultant curve is itself asymmetrical and falling to the right! **Moreover, for data or densities lacking that crucial fall to the right, such as the Uniform, the Normal and others, no matter how big the order of magnitude happened to be, nothing resembling the logarithmic is found. Here, order of magnitude does not play any role in logarithmic behavior! Uniform(0,  $10^{15}$ ) or Normal( $10^{19}$ ,  $10^4$ ) for example are not even remotely near any logarithmic configuration in spite of their huge order of magnitudes. In extreme generality, it is the confluence of big order of magnitude coupled with a falling density to the right that may lead to the logarithmic distribution under the right conditions. Surely  $k/x$  defined over (10, 100) does not require order of magnitude larger than one, but this is an exception.**

An important consequence of the fact that all logarithmic data sets come with a falling density curve to the right is that the average is (almost always) larger than the median for data sets obeying Benford's Law. It is important to realize though that positive skewness does not provide a warranty that the mean is larger than the median, since there are counter examples.

Negative values are rarer in real-life data sets than positives ones, yet they are part of the totality of everyday data. In reality there are two tails for the AGD Interpretation of Benford's Law: an immense one to the right for the positive numbers, and another relatively smaller and lower one to the left for the negative numbers. Both densities meet at the origin 0 but do not connect because their concentrations are quite different; hence a sharp discontinuity exists at the origin. Density of negative ones is much lower around the origin than the density of the positive ones. Little reflection is needed to realize that all the arguments and simulations above could apply to negative numbers just the same if performed separately, but the mixing of the two might prove highly problematic.

## FALLING DENSITY CURVE WITH A PARTICULAR STEEPNESS

---

Surely Benford's Law requires a very particular rate of fall in the density in the aggregate. If data is restricted to (10, 100), then that rate of fall is exactly dictated by the digital proportions of first, second, third, and higher orders. For example, exactly 30.1% of the area under the curve should belong to (10, 19.999...), 17.6% to (20, 29.999...), and only 4.6% to (90, 99.999...). Also within (10, 20) and within each 'decade', second-order law dictates a particular fall compatible with second-order digital proportions, hence more area should be allocated to (10, 10.999...) where second digit is 0, than to (19, 19.999...) where second digit is 9. All this, plus the consequences of the third, and higher orders laws, imply a very particular fall here, uniquely determined by the general law (which is in fact the rate of the fall of  $k/x$  distribution for this particular case of data falling between 10 and 100).

To better illustrate the connection between digital behavior and sharpness in the fall of the one-sided tail to the right, a table is included in Fig. 4.51 showing a variety of distributions, each with a different rate of fall. The table also helps in emphasizing the uniqueness of the distribution  $f(x) = k/x$ , and to differentiate it from all other distributions having a one-sided tail to the right of the form  $f(x) = k/x^m$ . The table assumes that all distributions are defined strictly on (10, 100), although the results obtained here are general and apply to all such distribution forms defined over intervals bordering other IPOT values, adjacent or not.

As can be deduced from the table of Fig. 4.51, only one particular sharpness in the fall of the tail to the right results in logarithmic behavior, namely that of the  $k/x$  distribution ( $m = 1$ ). Of interest here is the observation that related log density of  $k/\text{SQRT}(x)$  is shaped upward, that of  $k/x$  is flat, and that of  $k/x^2$  and higher powers are shaped downward.

The chart in Fig. 4.52 is the superimposition of all these six competing densities on the range of (10, 100). For better clarity, only the segment (10, 55) is shown. The thick line of  $k/x$  is the only one with the correct logarithmic rate of fall.

	0.073	0.434	11.1	202.0	3003.0	40004.0
	$\sqrt{X}$	$X$	$X^2$	$X^3$	$X^4$	$X^5$
Digit 1	19.2	30.1	55.6	75.8	87.6	93.8
Digit 2	14.7	17.6	18.5	14.0	8.8	5.0
Digit 3	12.4	12.5	9.3	4.9	2.1	0.8
Digit 4	10.9	9.7	5.6	2.3	0.8	0.2
Digit 5	9.9	7.9	3.7	1.2	0.3	0.1
Digit 6	9.1	6.7	2.6	0.7	0.2	0.0
Digit 7	8.4	5.8	2.0	0.5	0.1	0.0
Digit 8	7.9	5.1	1.5	0.3	0.1	0.0
Digit 9	7.5	4.6	1.2	0.2	0.0	0.0

Figure 4.51 Digits of Densities of the Form  $f(x) = k/X^m$  Defined over (10, 100)

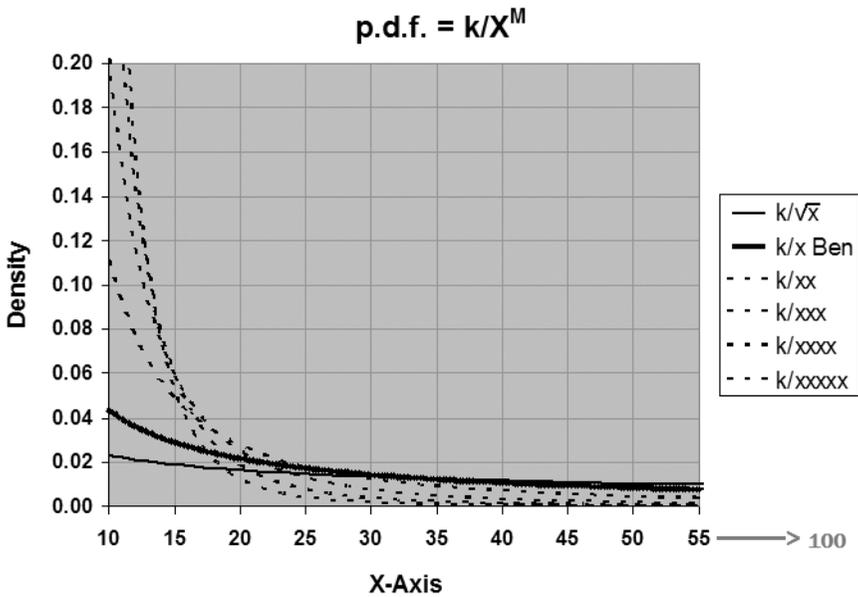


Figure 4.52 Superimposition of all Six  $k/X^m$  Densities on (10, 100)

## FALL IN DENSITY IS WELL-COORDINATED BETWEEN IPOT VALUES

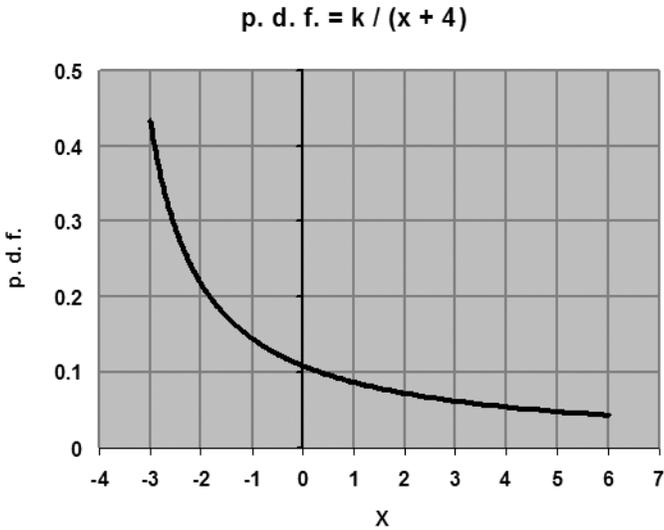
---

A density falling off as sharply as in the case of  $k/x$  is not a sufficient condition for logarithmic behavior. An essential criteria and a constant feature of logarithmic behavior is that such a fall must be properly aligned and coordinated with those relevant intervals between adjacent integral powers of ten. Acknowledging this fact is crucial for a solid understanding of the leading digit phenomena.

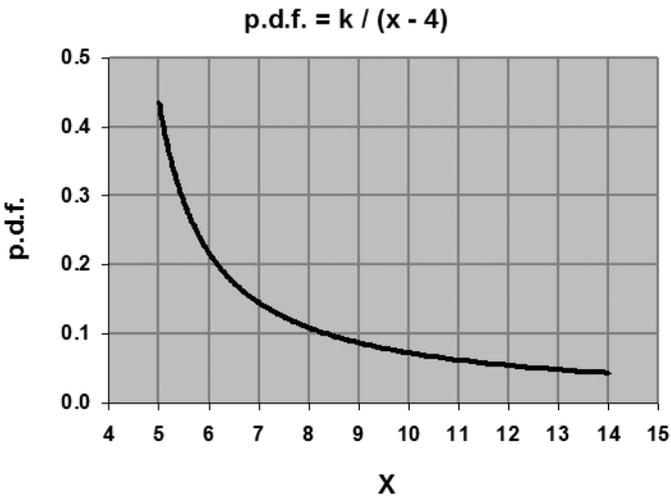
The following illustrative examples help to shed some light on this crucial issue. The first example is a translation of  $k/x$  defined on  $(1, 10)$  to the left by four units. The chart in Fig. 4.53 depicts one possible serious pitfall wherever logarithmically-proper density curve of data falls over the wrong interval. The distribution of the chart is one with the same shape as of  $k/x$  but it is defined over  $(-3, 6)$  instead of the more appropriate range of  $(1, 10)$ ; being shifted to the left by four units. Its density is  $\text{pdf} = k/(x + 4)$  defined over  $(-3, 6)$ . Here digits distribution is decidedly non-logarithmic; furthermore, it's not even monotonically decreasing, as digit 2 leads the most, gathering 39% of leadership.

First-digits distribution here is  $\{28\%, 39\%, 8\%, 8\%, 7\%, 2\%, 2\%, 3\%, 3\%\}$ , and digit 2 takes by far the most leadership. Such a distribution may represent, say, net profit in millions of dollars of a firm struggling not to stay in the red. The firm is having most of its probable values decidedly negative, some near zero, and very few positive ones. Note that there are actually two separate pitfalls in this example. The first is having the fall in such density shifted to the left, misaligning it over IPOT points and the normal digital order. The second is the perilous crossing of the origin (albeit in a smooth fashion), as if pretending that it doesn't exist, or that there is nothing special in passing such a game-changing point where the positive abruptly changes into the negative and vice versa.

A similar yet different pitfall is shown by the second example depicted in Fig. 4.54, where  $k/x$  is shifted to the right by four units, namely  $\text{pdf} = k/(x - 4)$  over  $(5, 14)$ , instead of the more appropriate one over  $(1, 10)$ . This shift does not involve any negative values whatsoever, yet its first digits do not resemble the



**Figure 4.53** An Adverse Shift in Focus and Mixing of Positive and Negative Values



**Figure 4.54** A Properly Shaped Curve is Translated — Distorting Digital Order

logarithmic in any way and it's not even monotonically decreasing. First digits are {22%, 0%, 0%, 0%, 30%, 18%, 12%, 10%, 8%}. Digit 5 draws the most leadership, gaining 30%. Digit 5 has the advantage that the density is being abruptly launched at 5.00. The density falls off from there steadily throughout hence the area on (5, 6) represents a large and significant portion of overall data.

In the limit, the curve  $\text{pdf} = k/(x - 4)$  exhibits logarithmic behavior farther to the right if defined over much longer ranges. This is so because it attains the proper shape of the  $k/x$  curve over proper intervals, and is in synch with IPOT values in the limit where 4 is quite negligible as compared with much larger values of  $x$ . For example, for  $\text{pdf} = k/(x - 4)$  defined over (5, 1004), the portion farther to the right on the interval (100, 1000) is extremely close to the logarithmic, but the part on the interval (10, 100) when examined in isolation is not logarithmic at all. The issue here of IPOT coordination is very general and does not only concern  $k/x$  distribution. The same adverse digital results seen in this chapter happen when, say, the Lognormal distribution with large enough shape parameter exhibiting logarithmic behavior is shifted to the right or to the left, as well as with any other logarithmic data or distribution curves.

In all such examples, we encounter distributions having the appropriate density shape for logarithmic behavior, falling at the correct logarithmic rate, yet badly oriented on the  $x$ -axis, falling over the wrong intervals. In other words, the densities are being out of phase with those intervals bordered by adjacent integral powers of ten and hence not logarithmic.

One then wonders how is it that real-life logarithmic data sets and especially AGD just happen to be so perfectly oriented on the  $x$ -axis so that the logarithmic is realized. This fortunate 'coincidence' of such perfect alignment appears a bit far-fetched and miraculous. A resolution of this paradox may be found in the claim that there is normally no need for any active and intentional orientation and that passively and indirectly data falls into such an orientation by default. To see why, let us first make a distinction between the following two types of logarithmic densities:

- (I) A density that starts out abruptly very high at a certain point on the  $x$ -axis and then keeps falling consistently from there having a steady tail to the right [typical for  $k/x$ ].
- (II) A density that first rises, hangs up there high about flat momentarily, and then reverses direction and consistently falls for a long duration [typical for Lognormal].

For type (I) data, most of the data (area under the curve) is near the starting point and it is not so easy to envision it as logarithmic unless the starting point is some IPOT value (perhaps 0, 1, or 10). Otherwise digit 1 would have to struggle hard to obtain its large 30.1% share. Such misplaced launching at 5.00 in Fig. 4.54 prevented digit 1 from earning its due logarithmic share of 30.1%. This is not a

hard and fast rule though, and Fig. 4.13 is a testament to this: a perfectly logarithmic distribution  $0.4342945/x$  defined over  $(30, 300)$  with 30.1% granted to digit 1 in spite of that strong launch at 30. It has managed to accomplish this by compensating digit 1 generously on  $(100, 200)$  for a long stretch. Note that what we call in this chapter the ‘starting point’ is what was called Lower Bound (LB) in all of the averaging schemes earlier. In summary, type (I) data does require often active anchoring to IPOT because its digital situation is delicate. For type (II) data, only a small portion of overall data is near the starting point. In fact much of it is in the center and farther to the right. For such type (II) density curve it does not matter much as far as digit configuration is concerned where the starting point is at; and it could definitely be at some non-IPOT value and yet quite logarithmic. In other words, for type (II) curves the logarithmic is concocted mostly around the central area where the bulk of the data resides between several IPOT points, and starting point is not very relevant. Typical pieces of everyday data are for the most part of type (II), and thus starting point is of no consequence mostly. For that smaller portion of type (I) real-life data that is logarithmic in its own right, it might be argued that at times the natural pull that the generic numbers 0 and 1 exert plays a crucial rule here as they are being utilized as anchors serving to coordinate and align the fall in the density with those intervals bordered by adjacent integral powers of ten values. Certainly 0 and 1 are the starting values of so much of our data: 0 for measurements, and 1 for count data. Fortunately for the logarithmic distribution in those rare cases as well, 0 and 1 just happened to be IPOT numbers! What a favorable coincidence!

For a real-life illustration of the enormous importance of 1 or 0 as anchors in digital configurations for densities of type (I), let us explore the adverse implication of their absence. Imagine a country where the king arbitrarily decrees that all streets must start with the number 5, being a holy or lucky number in their religion or folklore. We perform a Greek-Parable-like scheme to model this hypothetical street address data, similar to what was done in the simple averaging schemes, but with LB firmly fastened to 5. All streets have the sign at the door of the first house showing #5, and then #6, #7, #8, and so forth for subsequent houses. Street length varies from one to nine houses. We now allow in the model two-digit numbers such as 13, extending their number system slightly beyond nine.

All this is almost equivalent to a translation by four units to the right of whatever histogram was obtained in the original Greek Parable (applied to street data having a benevolent and reasonable king allowing for houses to start at 1) to arrive

| Street |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      |
| 5      | 5      | 5      | 5      | 5      | 5      | 5      | 5      | 5      |
|        | 6      | 6      | 6      | 6      | 6      | 6      | 6      | 6      |
|        |        | 7      | 7      | 7      | 7      | 7      | 7      | 7      |
|        |        |        | 8      | 8      | 8      | 8      | 8      | 8      |
|        |        |        |        | 9      | 9      | 9      | 9      | 9      |
|        |        |        |        |        | 10     | 10     | 10     | 10     |
|        |        |        |        |        |        | 11     | 11     | 11     |
|        |        |        |        |        |        |        | 12     | 12     |
|        |        |        |        |        |        |        |        | 13     |

Figure 4.55 Distortion in Digits Absent Those Ubiquitous Anchors 0 and 1

at the new histogram for the newly crowned eccentric ruler (constituting a similar transformation to that of  $k/x$  into  $k/(x - 4)$  seen in Fig. 4.54). For simplicity let us assume that not all streets get equal importance in terms of having the same number of houses, hence the shortest street represents only one house (showing the strange sign #5) while the longest street represents nine houses (showing signs from #5 to #13). This simplicity in calculations does not alter results by much as it points to the same conclusion almost. Figure 4.55 depicts the scheme.

First digits distribution for these 45 houses within the entire scheme is:  $\{10/45, 0/45, 0/45, 0/45, 9/45, 8/45, 7/45, 6/45, 5/45\}$  and calculated proportions are  $\{22\%, 0\%, 0\%, 0\%, 20\%, 18\%, 16\%, 13\%, 11\%\}$ . Hence digit 1 still squeezed a narrow and unexpected victory, but a lot of focus is given to digit 5, standing at the second place in leadership, simply because ‘density’ starts with this digit! A clearer result would have been obtained had we assigned equal importance to all streets and performed a bit more complicated calculations as was done in the original Greek Parable. Now with the assumption of equal street importance, digit 5 earns its deserved status as having by far the most leadership, and digits distribution is  $\{14\%, 0\%, 0\%, 0\%, 31\%, 21\%, 16\%, 11\%, 7\%\}$ . The similarity of either distribution above to the digital proportions of  $k/(x - 4)$  seen in Figure 4.54 is striking, and certainly expected. Without any benevolent intentions, the harsh and imposing king has provided us with a vivid flesh and blood narrative for this odd distribution. Clearly there is not even a resemblance to the logarithm for either proportion above, not even to Stigler’s Law. Absent here even that monotonically declining probabilities pattern. Such is the heavy price we must pay for blindly and

mindlessly obeying the arbitrary decrees of the king and losing those crucial and natural anchors of 1 and 0.

On a more profound level, the failures of  $k/(x + 4)$  defined over  $(-3, 6)$  and of  $k/(x - 4)$  defined over  $(5, 14)$  to behave logarithmically emanate from the fact that their densities are not exactly inversely proportional to  $x$ . Such an exact relationship between  $x$  and its density height is unique to  $k/x$  distribution, and it constitutes an essential feature of its logarithmic behavior. For any  $k/x$  distribution, doubling  $x$  causes the density to be cut (exactly) in half. For example, for  $0.4342945/x$  over  $(10, 100)$ , density height or histogram count on  $x = 40$  is exactly half the height or count on  $x = 20$ ; while the two non-logarithmic distributions  $k/(x + 4)$  and  $k/(x - 4)$  lack this property. In this sense, the preferred view about the latter two distributions is not that they are positioned in an uncoordinated way along the  $x$ -axis in the narrower context of digits, but rather that they are not exactly inversely proportional to  $x$  in the more general quantitative sense.

## SYNTHESIS BETWEEN THE DETERMINISTIC AND THE RANDOM

---

---

It may seem far-fetched that two totally different processes — deterministic multiplication processes on one hand, and probabilistic data, distributions, and processes on the other — are stamped with an identical logarithmic digital signature! What is there in common between the random final bill paid by a customer at a large supermarket (Random Linear Combination) and the steady and predictable increase in the monthly balance of an interest bearing account in a bank (Exponential Growth Series)? Yet the first digits in both themes are  $\text{LOG}(1+1/d)!$  There must be not only some minor digital commonality here but also another much more profound and intrinsic correspondence between the two.

Let us then select two representatives (one from each camp) and search for some fundamental common denominator: (1) The simplest of the simple averaging schemes, namely the Greek parable, representing a collection of random distributions, (2) The first 21 elements of 12% exponential growth series starting from base 1, representing deterministic multiplication processes.

In order to put both processes on an equal footing and enable comparisons, it is suggested here to view both of them as **static data sets**, not as dynamic processes. Hence any deterministic process is to avail itself also as a collection of numbers all with equal discrete probabilities. For example: the long deterministic (dynamic) process of obtaining all the prime numbers up to 50 could be viewed (at the end) as in picking randomly from the (static) set of elements  $\{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47\}$  all having equal probabilities. Hence,  $\text{Probability}(\text{Prime} < 11 \mid \text{primes up to } 50) = 4/15$ , and so forth. Money in a bank account dynamically earning interest for 50 years as in exponential growth series is to be viewed (at the end of the fiftieth year) as a static set of 50 values all equally likely. Such an approach would enable us to contemplate and construct histograms or density curves for both processes and make comparisons.

Moreover, the purely random distribution concept could be viewed as a data set as well simply by generating a fixed number of simulated values from it, and

then having the resultant data set representing it. For example, the Normal distribution with mean 7.5 and standard deviation 0.3 could be represented by {7.75, 8.04, 7.68, 7.59, 6.98, 7.74, 7.23, 7.24, 7.70, 7.45, 7.71, 7.15, 7.80, 7.15, 7.32, 7.87, 7.82, 7.74, 7.82, 8.09, 7.96, 7.33, 8.16, 7.08, 7.68, 7.45, 7.18, 7.27, 7.94, 8.04, 7.80, 7.34, 7.86, 7.19, 7.83, 7.38, 7.34, 7.66}. It must be noted that once these values of the Normal have been generated and presented, they are now considered all equally likely, as in the discrete Uniform distribution case.

Let us reconsider the simple averaging schemes, but instead of averaging the various digital distributions of the various individual intervals as was done earlier, we first view each interval as a mini data set, then combine all these various mini data sets into one single large data set, and finally measure digital pulse (only once) of that grand data set. Since each interval is accorded equal weight and importance within the model, and since intervals come with different lengths, it is necessary to 'stretch' them out by using repetitions until all are of equal length, namely the length of the longest interval. This is done to avoid giving longer intervals more weight than shorter intervals. In essence, we wish to treat the intervals as data sets and combine them all in one grand collection of numbers. The motivation here is to attempt to glance at the actual density form of the imaginary data itself (and its other properties perhaps) instead of just focusing on resultant digital distribution. For the Greek parable we assume that one out of nine topics will be about a spouse, one out of nine topics will be about houses, and so forth. To accord equal importance, we stretch all topics to nine values by assigning nine conversations per topic, resulting in a total of 81 typical numbers in typical conversations, namely [9 topics]\*[9 conversations]. The table in Fig. 4.56 represents one such attempt to arrive at a grand Greek data set of conversations.

Care should be taken in filling in the blanks for the numbers on the upper-right corner (shown in large and bold font). Surely we ought to imitate on the right whatever is on the left side of each row, row by row. We first simply repeat them in order, and then, when there is no more room for another full repetition of the entire set, we pick at random from the numbers on the left. Hence we note the slightly arbitrary manner in which the few remaining numbers on the extreme right side are chosen but which can be overlooked due to the fact that variations have very minor overall effect on results.

The grand data set for this particular attempt presented in the table of Fig. 4.56 is: {1,1,1,1,1,1,1,1,1,2,1,2,1,2,1,1,2,3,1,2,3,1,2,3,1,2,3,4,1,2,3,4,3,1,2,3,4,5,1,2,3,4,1,2,3,4,5,6,2,4,6,1,2,3,4,5,6,7,3,6,1,2,3,4,5,6,7,8,5,1,2,3,

<b>Spouse</b>	1	<b>1</b>							
<b>Houses</b>	1	2	<b>1</b>	<b>2</b>	<b>1</b>	<b>2</b>	<b>1</b>	<b>2</b>	<b>1</b>
<b>Slaves</b>	1	2	3	<b>1</b>	<b>2</b>	<b>3</b>	<b>1</b>	<b>2</b>	<b>3</b>
<b>Oranges</b>	1	2	3	4	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>3</b>
<b>Sheep</b>	1	2	3	4	5	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
<b>Chicken</b>	1	2	3	4	5	6	<b>2</b>	<b>4</b>	<b>6</b>
<b>Dogs</b>	1	2	3	4	5	6	7	<b>3</b>	<b>6</b>
<b>Olives</b>	1	2	3	4	5	6	7	8	<b>5</b>
<b>Gods</b>	1	2	3	4	5	6	7	8	9

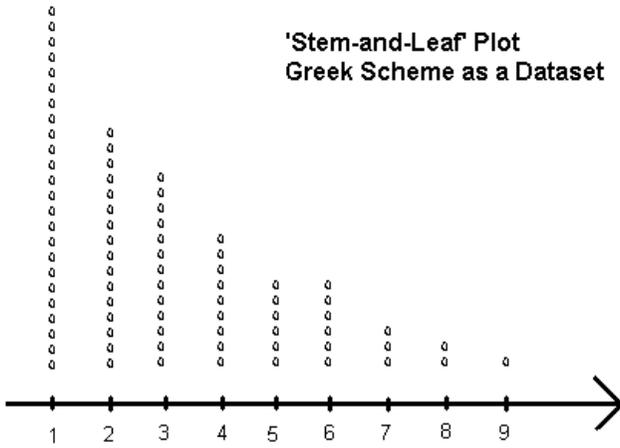
Figure 4.56 The Greek Leading Digits Parable as a Data Set

4,5,6,7,8,9}. The digit distribution of this grand data set is {30.9, 19.8, 16.0, 11.1, 7.4, 7.4, 3.7, 2.5, 1.2}. While not exactly equal to the result of the original Greek parable where digital configuration was obtained by averaging out the nine different mini digital configurations of each topic, nonetheless it is of course extremely close to it. The idea imbedded in the data set above, when extended and made more complex, could help in the visualization and the intuition of the scheme suggested by Flehinger and shown mathematically to equal  $\text{LOG}(1+1/d)$ .

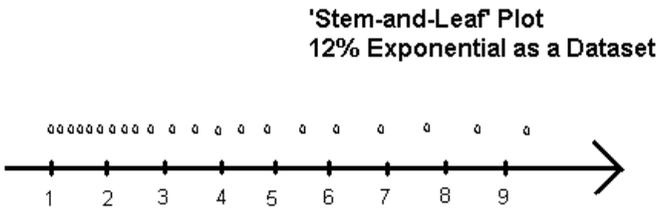
Treating the collection of the first 21 elements of 12% exponential growth series that starts at 1 as a data set we obtain: {1.00, 1.12, 1.25, 1.40, 1.57, 1.76, 1.97, 2.21, 2.48, 2.77, 3.11, 3.48, 3.90, 4.36, 4.89, 5.47, 6.13, 6.87, 7.69, 8.61, 9.65}. First-digits distribution for this short series is {33.3, 14.3, 14.3, 9.5, 4.8, 9.5, 4.8, 4.8, 4.8}. While digital distributions for these two deterministic and random processes are not equal, they are similar, and in the limit though both approach  $\text{LOG}(1+1/d)$ , converging to a singular and identical digit configuration!

Let us glance at these two very different data sets using a variation on stem and leaf plot, a histogram of sorts, as shown in Figs. 4.57 and 4.58.

The common denominator here can be easily visualized: it's simply the frequency per unit length, or density, which is diminishing for both data sets as focus shifts to the right. Both appear as having lopsided densities of sorts with tails to the right; both have numerous small quantities and very few big quantities. Yet, two points are still left unexplained and shrouded in mystery: (I) the fall in the densities for both is nicely aligned and synchronized between IPOT values in the limit, (II) the sharpness in the fall of the densities (steepness) is the same for both processes in the limit, coming out exactly as in the  $k/x$  case in the aggregate.



**Figure 4.57** The Random Process — Plotting the Greek Scheme



**Figure 4.58** The Deterministic Process — Plotting 12% Growth from 1

Such configuration of relative quantities as seen in Figs. 4.57 and 4.58 where the small is more beautiful than the big, is in harmony with Figs. 1.5 (B) and 1.5 (C) of the 10 by 10 multiplication table, where small quantities also outnumber big ones by a large margin, and which strongly corroborate the principle. This is not a coincidence, but rather another crucial source of logarithmic behavior in real-life data relating to the phenomenon of the multiplications of random measurements so frequently encountered in scientific and physical data sets, as shall be discussed in later chapters.

## DICHOTOMY BETWEEN THE DETERMINISTIC AND THE RANDOM

---

Turning our attention again to exponential growth series, we shall attempt to prove in general that distances between elements are expanding, as was seen in the previous chapter for 12% growth and visualized very clearly in Fig. 4.58. This in turn would prove that their 'discrete density' curve is falling. In the outline of Fig. 4.59, the constant value  $F$  stands for the multiplicative factor per period for a given percent growth, hence  $F = (1 + \text{percent}/100)$ . Since growth is assumed to be positive (as opposed to decay),  $F$  is always more than unity and the inequality  $F > 1$  holds. For example, for 6% growth the value of  $F$  is 1.06. For 100% growth, namely the doubling of quantities per period, the value of  $F$  is 2. The value  $B$  represents the base amount before any growth at time zero.

Let us now calculate distances between elements in exponential growth series as follows:

$$BF - B = B(F - 1)$$

$$BFF - BF = B(F - 1)F$$

$$BFFF - BFF = B(F - 1)FF$$

$$BFFFF - BFFF = B(F - 1)FFF$$

Clearly, distances between points are  $B(F - 1)F^N$ , where  $N$  stands for the index number of the  $N$ th element. Since  $F > 1$ , distances keep growing, as is visualized in Fig. 4.59. Hence, occurrences on the x-axis to the right become progressively less and less frequent as per any fixed length (i.e. less dense and more diluted to the right). Furthermore, examining the generic form of related log of exponential growth series yields important and surprising results and greatly aids in understanding their behavior. For exponential growth such as

$\{B, BF, BF^2, BF^3, \dots, BF^N\}$ , related log series is as follows:

$\{\text{LOG}(B), \text{LOG}(B) + \text{LOG}(F), \text{LOG}(B) + 2*\text{LOG}(F), \dots, \text{LOG}(B) + N*\text{LOG}(F)\}$ .

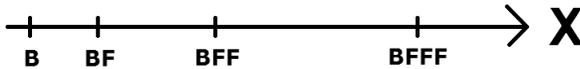


Figure 4.59 Exponential Growth Series on the X-Axis

Since this log sequence is simply the steady additions of the same constant, namely  $\text{LOG}(F)$ , from a fixed point, namely  $\text{LOG}(B)$ , the concentration per unit length is steady on the log-axis, and thus (discrete) ‘density’ is uniform on the log-axis! It is certainly proper to think of  $\text{LOG}(F)$  namely of  $\text{LOG}(1 + \text{percent}/100)$  as being a fraction as it is so in any case up to 900% growth. For example, for 2% growth,  $\text{LOG}(1 + 2/100) = 0.009$ . For 50% growth,  $\text{LOG}(1 + 50/100) = 0.176$ . Even for 180% growth,  $\text{LOG}(1 + 180/100) = 0.447$ , which is a fraction. In any case the above conclusion of uniformity of log does not require that  $\text{LOG}(F)$  is a fraction, although it being a fraction facilitates the visualization of log density being continuous-like and uniform. One must always keep in mind the essential feature of exponential growth series in the context of Benford’s Law, namely that the series is considered in its entirety, retaining all its intermediate products and discarding none, as opposed to considering, say, only the last element  $BF^N$ . We conclude that:

**Proposition V: Deterministic multiplication process of the exponential growth type is characterized by related log density being uniformly distributed (albeit discretely, not continuously).**

But by Proposition II, the approximate or most appropriate ‘density function’ here for the exponential growth series itself is  $k/x$  [the fact that related log is not continuous but rather discrete does not invalidate applying Proposition II to it]. Hence the following important conclusion is reached:

**Proposition VI: Deterministic multiplication process of the exponential growth type is of the  $k/x$  distribution albeit in a discrete fashion, having consistent and steady logarithmic behavior all along its entire range. The logarithmic requirement of an integral span on the log-axis (an integral exponent difference) applies here just as it does for any  $k/x$  distribution. The approximate [continuous] ‘density’ of the [discrete] exponential growth series is seen as falling steadily on the right at the constant logarithmic rate all throughout its entire range, mimicking  $k/x$ .**

**Philosophically and conceptually, this result demonstrates the intimate connection that exists between Benford's Law (*generic logarithmic rate of fall as in  $k/x$* ) and multiplication processes (*exponential growth series*), albeit as in multiplication processes of the particular deterministic exponential growth type and where all intermediate products are retained and considered.**

As a consequence of Proposition VI, for an exponential growth series with very large order of magnitude (i.e. large exponent difference) the requirement of Corollary I of an integral value of exponent difference can be easily waived, implying a near-perfect logarithmic behavior in all such long (log-wise) exponential growth series cases.

Exponential growth series are constructed with a constant multiplicative factor  $F_{\text{Constant}}$ . Another possibility to consider is that of the random multiplicative factor  $F_{\text{Random}}$ . For example, random factors constantly chosen from the Uniform on (1.23, 1.67) for each growth period could be utilized, leading to an exponential with haphazard and constantly changing growth rate (randomly fluctuating between 23% and 67%).

$\{B, BF, BF^2, BF^3, \dots, BF^N\}$  standard exponential growth series

$\{B, BU_1, BU_1U_2, BU_1U_2U_3, \dots, BU_1U_2U_3\dots U_N\}$  randomized exponential growth series

Such hybrid models fusing the deterministic and the random yield overall uniform related log distribution just the same, by the same line of reasoning above, and hence this random exponential series is logarithmic as well. This is so because its related log can be thought of simply as arising from slight random disturbances (displacements) of an original equidistance arrangement of values along the log-axis relating to compatible standard exponential growth series (one having a constant and predictable factor equals to the average of all these random  $U_N$ ). For a more decisive argument one can simply examine related log series of the random factors series:

$$\begin{aligned} &\{\text{LOG}(B), \\ &\text{LOG}(B) + \text{LOG}(U_1), \\ &\text{LOG}(B) + \text{LOG}(U_1) + \text{LOG}(U_2), \\ &\text{LOG}(B) + \text{LOG}(U_1) + \text{LOG}(U_2) + \text{LOG}(U_3), \\ &\dots \\ &\text{LOG}(B) + \text{LOG}(U_1) + \text{LOG}(U_2) + \text{LOG}(U_3) + \dots + \text{LOG}(U_N)\} \end{aligned}$$

Clearly its related log series is an additive random walk on the log-axis, implying an overall uniform flat ‘density’ on the log-axis albeit in a discrete manner. Conceptually therefore, such hybrid models relate more to the deterministic flavor and less to the random flavor, because association is ideally measured by that all-important related log density as per the prominent role it plays in the context of Benford’s Law.

The term ‘deterministic’ is typically associated with its literal dictionary meaning such as “an inevitable consequence of antecedent sufficient causes”, or “phenomenon that is causally determined by preceding events or natural laws”. Yet, **in the context of Benford’s Law and throughout the rest of the book, the term ‘deterministic’ would refer to any data set or distribution having its related log uniformly distributed**, such as the discrete fixed-rate exponential growth series of any length or growth rate, the random exponential growth series with varying factors, as well as (the uniquely continuous)  $k/x$  distribution defined over any type of range (logarithmic or otherwise). It must be noted though that references to  $k/x$  distribution are all about its random aspect, as are all references to other statistical distribution curves such as the Normal, the Uniform, the Lognormal, and so forth. The distribution  $k/x$  in a purely statistical context is certainly random. In addition, besides encompassing discrete exponential growth series, the term ‘deterministic’ here would also refer to any discrete set of numbers such that approximately at least equal number of values fall between each sub-interval standing between adjacent IPOT points (i.e. steady dispersion between the integers of related log-axis). Examples of processes when literal dictionary meaning may apply can be found when exact parameters are specified for exponential growth series, such as a bank account frozen for, say, 30 years from January 1, 2011 to December 31, 2040, cumulatively growing at a fixed annual 5% interest rate paid on the last day of each December. Here nothing is left as random, and one can predict the balance in the account on January 1, 2041 with absolute certainty. The same literal meaning applies to dynamical system of the form  $S_{n+1} = S_n^2 + 1$ , the Fibonacci series, the factorial sequence  $\{N!\} = \{1!, 2!, 3!, 4!, \dots\}$ , the self-powered sequence  $\{N^N\} = \{1^1, 2^2, 3^3, 4^4, \dots\}$ , as well as to the set of prime numbers up to  $N$ , to mention just a few cases. Yet, to call any one of these dynamical systems or sets of numbers ‘deterministic’ in our context of BL, one must first verify that related log is indeed uniform, albeit only approximately so or only in a discrete sense.

It has often been (wrongly) stated that the logarithmic distribution can be explained if one assumes that numbers are just as likely to be (there are just as many numbers) between say 10 and 100 (logarithm between 1 and 2) as between say 100 and 1000 (logarithm between 2 and 3), and so forth; namely uniform and equal concentrations of numbers across intervals standing between adjacent IPOT numbers. It is easy to see how this assertion leads to the statement that everyday numbers become more sparse and rarer on higher ranges of the natural numbers, since having a fixed number of values say 40,000 on (10, 100), coupled with having the same number of 40,000 values also on the much longer interval (100, 1000), necessitates that values are much more diluted on the latter (100, 1000) interval in comparison. This lowering of concentration (density) of numbers along the natural number axis then points to an asymmetric one-sided tail to the right.

The above assertion is conceptually and vaguely true, but not exactly in the manner thus stated. The assertion goes (unnecessarily) one step beyond Newcomb's statement, implying that not only distribution of **mantissa is uniform**, but that **log itself is uniform** as well. Yet, typical everyday data do not come with a flat related log density at all, and concentration between adjacent IPOT values almost always builds up gradually on the left portion where low values reside, and then diminishes and dies out gradually on the right where high values reside. Instead of maintaining a steady count within those sub-intervals standing between adjacent IPOT points, typical real-life data builds it up on the left and diminishes that count on the right. The above assertion is only true for exponential growth such as interest-bearing account in a bank that is frozen for many years or decades, many weeks of bacterial growth in the laboratory, and for that abstract 'purely logarithmic' distribution  $k/x$ . Related log densities has a way of revealing much about the data that neither density of data itself nor digital distribution can show, and therein lies the fundamental difference between the 'random' and the 'deterministic'.

**Proposition VII:** Empirical evidences strongly indicate that related log of almost all real-life data sets (what is typically called 'random', with the exception of exponential-growth-like multiplication processes) starts out from the very bottom of the log-axis, rises up gradually to some plateau, then falls gradually as well all the way back onto the log-axis, ending there, imitating slightly at times the shape of the Normal or the contorted semi-circular in some way. There are almost no exceptions to this observed universal pattern.

When the range on the log-axis is wide enough, the data has a near-perfect logarithmic behavior if considered in the aggregate over its entire range, but not over small segments within it. Deterministic multiplication processes such as exponential growth series on the other hand come with related log density that is uniformly distributed, always hanging above log-axis horizontally, steadily, and flat, having logarithmic digital behavior that is completely steady and consistent throughout. In extreme generality: the density of random data itself appears a bit similar to the Lognormal while the density of deterministic data itself is of the  $k/x$ -like type.

**Applicability of  $k/x$  Distribution: In spite of its perfect logarithmic behavior and its uniqueness whenever range is restricted to two adjacent IPOT points (being indeed the definition of the Benford condition there), the distribution  $k/x$  is NOT the appropriate model for typical everyday data sets Benford's Law mostly refers to. The distribution  $k/x$  is exclusively connected with deterministic multiplication processes best exemplified by exponential growth series of the form  $\{B, BF, BF^2, BF^3, \dots, BF^N\}$ , all of which are quite rare in real-life data sets. In conclusion,  $k/x$  distribution has very limited relevance to real-life empirical application of Benford's Law. Theoretically, though,  $k/x$  plays a crucial role.**

An indirect consequence of the dichotomy between the Random and the Deterministic is that small slices of large logarithmic data sets exhibit distinct digital behavior, depending on whether type is random or deterministic. In general, any subset of sufficient size cut out from some large logarithmic data set is still logarithmic, given that extraction from the whole is truly random, or equivalently, whenever the whole has been well mixed itself before an ordered extraction on the newly shuffled data set took place. On the other hand, when a logarithmic data set is ordered (i.e. sorted low to high say), cutting off pieces and slices from it in an orderly fashion from particular sub-ranges will not result in logarithmic behavior at all unless logarithmic data is of the deterministic type and enough elements are included. For this reason, sequential segments of exponential growths or decay series (i.e. deterministic) focused on a particular sub-range could still inherit the logarithmic property of the whole, even though the series is being sampled with its order fully intact. On the other hand, focused subsets from random data are not logarithmic at all, unless extraction is such that values are taken randomly from the whole range without any focus on specific sub-intervals.

An alternative point of view about the remarkable correspondence between the two disparate processes of the random and the deterministic can be given via the perspective of mantissa. Both processes arrive at uniformity of mantissa due to the very particular form of their related log densities: (I) the deterministic process achieving uniformity of mantissa directly by way of related log itself being flat and uniform over an integral range, (II) the random process achieving uniformity of mantissa by having its related log rising and falling gradually on the log-axis over large enough a range and without discontinuities or large coordinated dents, and so forth, as postulated in the statement of related log conjecture.

The importance of distinguishing between the random and the deterministic in the context of Benford's Law should not be overlooked. An acknowledgement of the dichotomy between the two processes helps in shedding light on many of its aspects. From its very inception, the field has been suffering from a profound confusion and mixing of these two very different logarithmic flavors within the discipline, causing mistaken factual conclusions and theoretical misconceptions. One example of such perilous omissions is the mistaken liberal application of Allaart's sum-invariance characterization of Benford's Law, namely equality of sums of actual quantities along digital lines, to whatever logarithmic data flavor or type under consideration. His characterization can only be applied to deterministic processes and, of those, only to one very particular class where the range stands exactly between two IPOT values, adjacent or not. Random data sets lack this sum equality altogether and consistently show significantly larger sums for lower digits. For example, the sum of all numbers beginning with digit 1 was thought to equal the sum of all numbers beginning with digit 2, and so forth. The misguided intuition here claims that summing those very few numbers beginning with 9 (of supposedly larger values), should yield something similar if not outright equal to summing all those much more numerous numbers beginning with digit 1 (of supposedly lower values), due to some grand imaginary trade-off. Pieter Allaart published in 1997 a rigorous proof showing equality. In Kossovsky (2006) it is noted that his proof is based on a restricted range spanning only two adjacent integral powers of ten (1, 10) which, according to Proposition III, implies the limited  $k/x$  case of the deterministic flavor. In communicating with Allaart he has readily accepted this qualification and acknowledged the limitation in application. Is it extremely rare that real-life data should fall into such restricted ranges, but even if it does, such data is by default of the deterministic flavor representing exponential growth series, not of the typical random flavor. By basing

his proof on such a restricted interval, indirectly or passively, Allaart chose to consider only the deterministic flavor having particular range standing between adjacent IPOT numbers. Empirical results from numerous real-life random data sets, and without a single exception, consistently confirm this theoretical conclusion, as sum earned by the lowest digit 1 is roughly five to even ten times the sum earned by the highest digit 9. In addition, computer simulations of the Lognormal and Exponential distributions, two densities closely related to numerous real-life random data types, consistently show an almost identical large disparity between the sums earned by digit 1 and digit 9. On the other hand, computer simulations nicely confirm Allaart's sum invariance characterization in all examined exponential growth series standing between IPOT points. To top it all off, computer simulations of  $k/x$  distributions with range standing between any IPOT points, strongly confirm Allaart's sum-invariance characterization, and equality along digital line are almost exactly observed. This last result perfectly fits the argument here restricting Allaart's scope and also corroborates his theoretical result with the purest form of the deterministic flavor of Benfordness. A later chapter in this section is devoted to Allaart's sum invariance characterization, containing detailed empirical and theoretical results as well as an alternative mathematical proof of such invariance.

The widespread incorporation and use in the literature of algebraic expressions such as  $S_{N+1} = F \cdot S_N$  or, as in many other typical expressions of exponential growth series such as  $S_N = B \cdot F^N$ , regarding proofs and general representation of data, **should not be considered appropriate** unless the researcher is dealing specifically with deterministic exponential growth or such (which are extremely rare in typical real-life data sets). It is erroneous to start a formal mathematical proof with such liberal use of these expressions since they are applicable only in the restricted case of the deterministic flavor. Moreover, utmost care should be exercised in deciding a priori whether or not proofs, setups, results, and conclusions depend on a restricted range such as IPOT end-points, integral exponent difference, unity exponent difference, or any other conditions on boundary. Yet, such algebraic representations as  $S_{N+1} = F \cdot S_N$  and  $S_N = B \cdot F^N$  are indeed appropriate when dealing in cases like hourly bacteria growth in the lab, undisturbed for many days; money balances in a frozen bank account, growing at a fixed interest rate, and considered over many years or decades without other interrupting transactions; population readings over many years of a single metropolitan (not countrywide aggregations) which does not

experience migration or devastating plague, and so forth. Clearly, all these cases are rare indeed.

Another important factor which has caused a great deal of confusion and misguided mixing in the literature is the apparent (but mistaken) similarity between MCLT-resultant Lognormal distribution and exponential growth series. Both relate to multiplicative processes but in a very different manner. The very fact that these two different manifestations/flavors of the logarithmic phenomena point to the same concept of the multiplicative process has misled many a scholar, deeming both of them equal in digital sense when in fact they are not! One should not speak of ‘multiplicative process’ per se, but rather must specify carefully the type of process involved.

There are four intrinsic differences that distinguish between the above-mentioned types. For **exponential growth**: (I) the multiplicative factor is constant/ fixed, (II) its value is thus deterministic, (III) the series is a single very long batch of products, (IV) all intermediate products are retained and considered as part of the whole set. For the **Lognormal and MCLT**: (I) the multiplicative factor varies constantly, (II) its value is random arising from a singular random distribution, (III) numerous batches of products having relatively few factors each are bundled together, (IV) all intermediate products within each batch itself are discarded, not to be considered as part of the set in question.

A data set having R elements (R realizations) of the Lognormal distribution originating from N multiplicative random processes of independent and identically distributed random variables X via the Mutiplicative Central Limit Theorem can be expressed as follows:

$$\{x_{11} * x_{12} * x_{13} * \dots * x_{1N}, \\ x_{21} * x_{22} * x_{23} * \dots * x_{2N}, \\ x_{31} * x_{32} * x_{33} * \dots * x_{3N}, \\ \dots, \\ x_{R1} * x_{R2} * x_{R3} * \dots * x_{RN}\}$$

with a fixed constant N that is not necessarily very large, and where  $X_{ij}$  signifies a distinct realization from X. The fact that intermediate products are discarded, not to be considered at all, implies that ‘partial products’ such as  $x_{11} * x_{12}$  or  $x_{11} * x_{12} * x_{13}$  are not part of the data.

In sharp contrast, data set generated by exponential growth series is constructed in a very different style of accumulation and bundling of products, and it

is expressed as:  $\{B, BF, BFF, BFFF, \dots, BF^N\}$ , where B is the base, F is the constant multiplicative factor, and with the process having a rather very large value of N sequences for logarithmic behavior to be expected. Although these two very distinct processes lead to the same digital result, namely the logarithmic distribution, they are not entirely digitally equivalent and their related log densities show a marked differentiation: flat in the exponential series case, and curved in the Lognormal case.

A third source of confusion appears to be the idea of a random selection involving a small sample from a large population of deterministic data, such as exponential bacterial growth or an interest-bearing bank account. If sample drawing is done in a truly random fashion then its related log density is just as uniform as population's related log density, and therefore its manifested nature is decidedly deterministic in spite of the random feature associated with the process! This is a bit tricky, and the statistician or forensic data analyst attempting to detect possible fraud via digital development examination should be very clear about the appropriate type and flavor associated with the data under consideration.

In conclusion: not all logarithmic data are created equal, but rather are created along two unique flavors: the consistent deterministic type, and the variable random one.

## FITTING THE RANDOM INTO THE DETERMINISTIC

---

---

The assertion that the prevalence of the logarithmic property in typical everyday data springs from the purported fact that most real-life data sets emanate from exponential-growth-series-like multiplicative processes must be rejected. Such erroneous assertion ultimately claims that the logarithmic has but one manifestation and a single unique flavor. It does not! It has been demonstrated earlier that there are two distinct and very different Benford types, the deterministic and the random. Moreover, the deterministic flavor is of lesser importance, constituting only a tiny fraction of typical real-life data. The vast majority of data we deal with is random. The assertion has its origin in Frank Benford 1938 seminal paper, philosophically stating that: **“We are so accustomed to labeling things 1, 2, 3, 4, ... and then saying they are in natural order that the idea of 1, 2, 4, 8, ... being a more natural arrangement is not easily accepted. Yet it is in this latter manner that a surprisingly large number of phenomena occur, and the evidence for this is available to everyone”**. While the eloquence in his prose is quite striking and his broad conceptual account describing the usage of our number system in relationship to the physical world is overall correct, they aren't exactly as in 1, 2, 4, 8 progression. Almost certainly what Benford had in mind is the tendency of natural phenomena to congregate in the regions of small values, and to become sparser and diluted in the regions of big values, as in a falling density curve with a one-sided tail to the right; that in nature the small outnumber the big. Such a tendency favoring the small is also perfectly manifested in exponential growth series as seen earlier, a fact which tempted and lured Benford to narrowly describe his law as such.

To disprove Benford's assertion, at least in an empirical sense, an attempt will be made to fit one random real-life data set into whatever exponential growth series that might be compatible with it. The attempt will prove futile, thus demonstrating that the assertion is wrong at least for the particular data set chosen, and also in general by induction (as well as by empirical testing of some other random data sets). Data on concentration of U.S. population in cities and towns in 2009

shall be selected for that purpose for three reasons: (I) it is very close to being almost perfectly logarithmic, (II) it is large enough in terms of size for a good analysis (with 19509 data points), (III) the flavor of all population data sets is unmistakably random, and this fact shall be demonstrated in the next chapter.

A failure to fit this population data into the mold of an exponential growth series should strongly argue against Benford's assertion. The existence of a profound difference between deterministic and random processes has already been demonstrated via the distinct shapes of their log densities, but here we go a step further by demonstrating it directly from the data itself, examining the different ways and distinct speeds actual numbers increase from low regions and into higher ones.

We shall order the entire data set low to high and a multiplicative factor would then be calculated between each pair of consecutive values, namely the factors as in  $X_{N+1} = F_N * X_N$ . The factor  $F$  is equal to 1 if there is no progress and the value repeats. The factor  $F$  is larger than 1 if there is even the slightest increase.

Out of 19,509 population values, 19,508 factors (1 less) were calculated. Out of these 19,508 factors, 11,558 were exactly 1, namely 59.2% of all cities had values that did not rise at all, but rather were repeated! This is certainly not a behavior consistent with the steady growth of an exponential growth series predicted by Benford! Interestingly, the spread of those non-progressive factors of 1 were not uniform across the data, but rather it was heavily biased towards the low values on the left in the beginning, dying out gradually towards the end on the right. Another strong dissimilarity with exponential growth series, which by definition comes with a single consistent factor, was the fact that progressive factors were for the most part very diverse. Out of 7,950 non-unity progressive factors, the vast majority of them — 6,763 to be exact — were totally unique factors (i.e. distinct), 1,032 were duplicates, 147 triples, and 8 quadruples. The table in Fig. 4.60 shows in more detail what is typically occurring within the data. As can be seen from the table, there are frequent and long repetitions of values in the beginning on the left, less so in the center, and none towards the end on the right. Also it is noted that non-unity progressive factors are almost consistently decreasing, from around 1.05 in the beginning, to around 1.0005 towards the end. This graduation in the values of the factors is consistent with values on the left being of low magnitude of course, and values on the right being of high magnitude, so that a left-transition from, say, 17 to 18 (an increase by just 1 unit) is much more meaningful (percent-wise) than a right-transition from, say, 9336 to 9340 (a larger increase by 4 units), where factors here are 1.0588 and 1.0004, respectively.

<i>Left Region</i>		<i>Central Region</i>		<i>Right Region</i>	
<u>Population sorted</u>	<u>Factor</u>	<u>Population sorted</u>	<u>Factor</u>	<u>Population sorted</u>	<u>Factor</u>
17	1	696	1	9336	1.0003
17	1	697	1.001437	9340	1.0004
17	1	697	1	9344	1.0004
17	1	697	1	9356	1.0013
17	1	697	1	9361	1.0005
17	1	697	1	9363	1.0002
18	1.0588	697	1	9365	1.0002
18	1	697	1	9373	1.0009
18	1	698	1.001435	9375	1.0002
18	1	698	1	9386	1.0012
18	1	698	1	9387	1.0001
18	1	698	1	9392	1.0005
18	1	699	1.001433	9396	1.0004
19	1.0556	699	1	9402	1.0006
19	1	699	1	9406	1.0004
19	1	699	1	9424	1.0019
19	1	699	1	9433	1.0010
19	1	700	1.001431	9435	1.0002
19	1	700	1	9436	1.0001
19	1	700	1	9439	1.0003
19	1	700	1	9453	1.0015
19	1	701	1.001429	9460	1.0007
19	1	701	1	9464	1.0004
20	1.0526	701	1	9465	1.0001

**Figure 4.60** Progressive and Non-Progressive Factors of U.S. Population Data

In order to gain additional understanding about the structure of the factors, let us examine factors in the generic case of the Lognormal distribution, under the assumption that it is a good representative of random processes and data in general. Simulation of the Lognormal with location parameter of 4.7 and shape parameter of 1.8 yielded 35000 realized values. The generated data was then sorted and factors obtained. The table in Fig. 4.61 shows three samples of factor structure from the left, center, and right regions; including the average factor for each region which is shown on top. Clearly almost all factors are non-unity progressive ones. The feature that seems the most stable and generic here is the U-shape-like of the average factors on a region-by-region basis. The chart in Fig. 4.62 depicts factor-averages from regions all over the range and the pattern seen here is very general, true for other parametrical values of the Lognormal so long as shape is high enough (approximately over 1) and distribution is logarithmic. In conclusion: it is not possible to fit the Lognormal into anything resembling exponential growth series since almost all the factors are distinct and vary from region to region.

Left region 1.0005		Middle region 1.0001		Right region 1.0007	
LogN sorted	Factor	LogN sorted	Factor	LogN sorted	Factor
7.9429	1.00121	96.7933	1.00003	3010.5207	1.00104
7.9456	1.00033	96.7993	1.00006	3012.0108	1.00049
7.9468	1.00015	96.8051	1.00006	3013.2761	1.00042
7.9514	1.00059	96.8213	1.00017	3017.4359	1.00138
7.9515	1.00002	96.8419	1.00021	3023.0057	1.00185
7.9532	1.00021	96.8422	1.00000	3030.2708	1.00240
7.9568	1.00045	96.8520	1.00010	3037.2250	1.00229
7.9605	1.00045	96.8682	1.00017	3037.6928	1.00015
7.9681	1.00095	96.8756	1.00008	3038.5468	1.00028
7.9731	1.00063	96.8819	1.00007	3040.5081	1.00065
7.9785	1.00068	96.8829	1.00001	3043.2008	1.00089
7.9787	1.00003	96.9248	1.00043	3044.9960	1.00059
7.9812	1.00031	96.9328	1.00008	3045.0158	1.00001
7.9844	1.00040	96.9463	1.00014	3045.3382	1.00011
7.9939	1.00119	96.9772	1.00032	3045.8869	1.00018
7.9945	1.00007	96.9996	1.00023	3047.3058	1.00047
7.9957	1.00015	97.0003	1.00001	3049.4485	1.00070
8.0023	1.00082	97.0115	1.00012	3049.5121	1.00002
8.0169	1.00182	97.0219	1.00011	3051.1809	1.00055
8.0198	1.00036	97.0234	1.00002	3053.9451	1.00091
8.0206	1.00011	97.0467	1.00024	3054.0573	1.00004
8.0215	1.00011	97.0544	1.00008	3055.4612	1.00046
8.0234	1.00023	97.0549	1.00000	3056.4383	1.00032
8.0271	1.00047	97.0563	1.00001	3060.7004	1.00139

Figure 4.61 Factor Structure of Lognormal ( $L = 4.7, S = 1.8$ ) for Three Regions

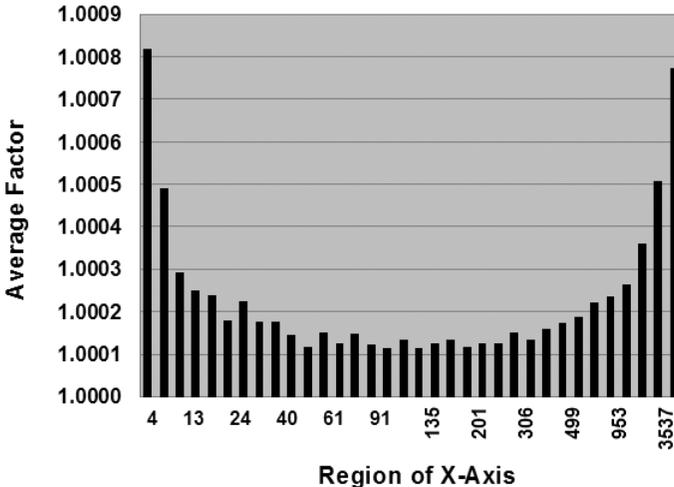


Figure 4.62 Factor Structure of Lognormal ( $L = 4.7, S = 1.8$ ) for Multiple Regions

Finally, in the same vein, two additional attempts were made to try to fit random data types into deterministic exponential growth series. Both ended in failure. The first test was performed for the empirical gathering of Internet data from a large variety of sources in the spirit of Hill's phenomenon of all phenomena of Chapter 110. The second test was performed for a chain of seven Uniform distributions. In conclusion, typical real-life random data does not occur as in 1, 2, 4, 8 even approximately. Rather, it occurs with a much more complex structure.

## THE RANDOM FLAVOR OF POPULATION DATA

---

Data on population within artificial or arbitrary (legal) divisions or boundaries, such as counties, cantons, prefectures, regions, and districts (each containing multiple concentration points of cities and towns) are in principle almost perfectly logarithmic just as populations of congregation points are, but due to their small data size they are usually not logarithmic enough. Counts of population centers for any large country on the other hand are nearly perfectly logarithmic. Some in the literature have argued that data on concentrations of population shows such strong agreement with the logarithmic because population supposedly grows exponentially at a constant or similar rate, and that census population data of numerous cities at one instance in time, namely a snapshot of different towns and cities at their different stages of developments is akin to multiple yearly snapshots of a single city growing over years or centuries. Omitted in such misguided model is the fact that one population center often draws from another, and such random migration from one small town to another big city (or vice versa) totally distorts any similarity to exponential growth series. The model to consider is deterministic exponential growth series expressed as  $\{B, BF, BFF, BFFF, \dots\}$  being transformed into a dynamic and random process of the form  $\{B, BF + I - E, (BF + I - E)F + I - E, ((BF + I - E)F + I - E)F + I - E, \dots\}$ , where  $I$  and  $E$  are random variables representing immigration and emigration, respectively. Certainly this model (constructed for just one city) does not resemble exponential growth series in any way.

Effective population growth in any single city is for the most part made of a few separate rates, such as death rate, birth rate and net migration rate, thus overall growth rate is certainly a random factor, not constant, especially when considered for numerous cities. The trajectory of the population of New York City from its very inception in 1624 to the present with all its twists and turns (ignoring for simplicity the prior settlements of the native Lenape people) could not even average anywhere near the snapshot of the current assortment of cities and towns in the USA that were mostly established around the years 1700–1900. Any current snapshot of the census office, for example, omits numerous past and futuristic

towns and cities in stages of development very different than the ones being recorded presently, unless one absurdly assumes that all the cities have been established gradually and steadily, say, one city per year, so that their ages are evenly spread. In any case, the worst fault in such erroneous model lies at the heart of the idea, since it preposterously claims that there exists a singular and equivalent deterministic exponential growth series perfectly fitting the set of all the last elements of the following random set of (supposedly nice and steady) exponential growth series:

- {138, 138\*F, 138\*FF, 138\*FFF, ... , **138\*F<sup>210</sup>**} — 210 years, Chicago, IL
- {203, 203\*F, 203\*FF, 203\*FFF, ... , **203\*F<sup>312</sup>**} — 312 years, Detroit, MI
- {27, 27\*F, 27\*FF, 27\*FFF, ..... , **27\*F<sup>166</sup>**} — 166 years, Salt Lake City, UT
- {9, 9\*F, 9\*FF, 9\*FFF, ..... , **9\*F<sup>349</sup>**} — 349 years, Albany, NY
- {5, 5\*F, 5\*FF, 5\*FFF, ..... , **5\*F<sup>102</sup>**} — 102 years, Las Vegas, NV

and so forth, for all the cities and towns in the USA. For example, Chicago was established in 1803 [210 years ago]. A wild and unruly band of 138 adventurous people on wagons is imagined to constitute the initial population size, and nobody has ever immigrated there to become Chicagoan! Moreover, should the model assume a common and constant F for all the cities, and that it never varies over time? In any case the model purports that the set {**138\*F<sup>210</sup>, 203\*F<sup>312</sup>, 27\*F<sup>166</sup>, 9\*F<sup>349</sup>, 5\*F<sup>102</sup>, etc. for all cities**} can be approximated somehow by a singular exponential growth series!?

The flavor of all population data sets is unmistakably random, not deterministic, and they are logarithmic not because they mimic any supposed exponential growth series, but rather because they are just one more example of the single-issue physical manifestation of Benford’s Law, like data on river flow, exoplanet mass, pulsar radiation rates, earthquake depth below the ground, brightness of astronomical objects, and so forth. Alas, the most decisive demonstration that population data has the random flavor is given by the examination of its related log distribution, which curves around nicely mimicking the Normal curve a bit, and does not resemble the flat uniform in any way! In line with one of the conjectures about single-issue physical phenomenon that will be given in later chapters, random multiplicative process may serve perhaps as the model for population data, and consequently, by way of Multiplicative Central Limit Theorem, the Lognormal should be its approximating density. Let P<sub>0</sub> stands for the initial population count

during establishment, and  $F_1, F_2, F_3, F_4,$  etc. be the yearly random and variable factors expanding or contracting the population count, then  $P_{\text{TODAY}} = P_0 * F_1 * F_2 * F_3 * F_4,$  and so forth. Here,  $F_N$  represents the variable growth factor which incorporates birth, death, and net migration rates, and which also adjusts for plagues, wars, natural disasters, and so forth. Another possible explanation for the logarithmic behavior of population data may be given by the model of random throws of balls into boxes to be discussed in a later chapter.

Kenneth Ross (2011) has also wrestled with the question of why population data sets so closely agree with Benford's Law, and has recently presented a rigorous mathematical proof showing that a large collection of the last term of exponential growth series  $BF^N$  where  $B$  and  $F$  are randomly selected from suitable uniform distributions is logarithmic in the limit as  $N$  goes to infinity. His model is of a census population snapshot of a large collection of relatively older and well established cites, all being established simultaneously in the same year, and with inter-migration excluded. Moreover, he provided a second model that allows for  $F$  to be randomly selected anew each period, and thus constituting a somewhat more realistic model of cities where growth rate is typically not constant over long periods. Monte Carlo computer simulations for the first model strongly confirm his assertion and give highly satisfactory results even after merely 20 growth periods (years). Computer simulations for his second model yield even faster convergence, with satisfactory results gotten even after only, say, 5 growth periods. Little reflection is needed to realize that his second model can be interpreted as repeated multiplications of random (uniform) distributions, which is Benford as predicated by the Multiplicative Central Limit Theorem (to be discussed in Chapter 79).

## THE LOGNORMAL DISTRIBUTION AND BENFORD'S LAW

---

The Lognormal distribution often represents a variable obtained from a product of many independent and identically distributed random factors, and thus it is commonly modeled on the Multiplicative Central Limit Theorem (MCLT), a prevalent and very important driving force in mathematical statistics. This explains why the Lognormal is so relevant and common in real-life data. The Lognormal has existence also outside repeated multiplications; numerous data sets and variables are Lognormal without any multiplicative origin whatsoever, and this fact renders the Lognormal even more prevalent and important in real-life data. It must be emphasized that a process of repeated multiplication is necessarily Lognormal, but the Lognormal is not necessarily a distribution representing a process of repeated multiplication.

Solely being Lognormal random variable  $Y$  via repeated multiplications of many independent and identical realizations of a random variable  $X$  is not a sufficient condition guaranteeing logarithmic behavior; it all depends on the spread of  $X$  (its standard deviation) as well as on how many times  $X$  is being multiplied. The larger the spread of  $X$  and the more we multiply, the closer we get to the logarithmic. In the limit when the number of repetitions goes to infinity the logarithmic is encountered even when the spread of  $X$  is small. Obviously, the digital analyst is not interested in the spread of  $X$  per se. Rather, he or she is simply seeking a good spread of  $Y$  having a large order of magnitude or equivalently a large range on its related log axis. Typically for real-life data sets, the spread of the generating  $X$  variable is large enough so that even with a few repetitions  $Y$  is sufficiently close to the logarithmic.

The distribution can be represented as **Lognormal** =  $X_1 * X_2 * X_3 * \dots * X_N$ , where the terms  $X_i$  refer to independent and identical realizations from  $X$ , and  $N$  in a strict sense is  $\infty$ , or a very large number in which case this refers to an infinite process of multiplication which converges to the Lognormal as  $N$  grows. In practical terms,  $N$  does not have to be very large at all in order to observe something

extremely close to the Lognormal. This remarkable result, namely that the actual distribution form of X or the values of its parameters are all immaterial as it can take on any form and utilize any parameter is derived from the MCLT, and it renders the result a truly large scope. It must be noted that all intermediate products are not being retained as in exponential growth. Rather, they are all being discarded at the end except the last one. Hence the Lognormal is **NOT** as in  $\{X_1, X_1 * X_2, X_1 * X_2 * X_3, X_1 * X_2 * X_3 * X_4, \dots, X_1 * X_2 * X_3 * X_4 * \dots * X_N\}$ , but rather only as in the last term.

The Lognormal should not be confused with what was noted earlier in Chapter 58 regarding raising a random variable to an integral power, namely  $X_1^N$ , where  $X_1$  refers to a single realization from any random variable X. The process  $X_1^N$  is also logarithmic in the limit as N gets large. In both cases a new variable is being concocted in a multiplicative manner from a single variable X, but the similarities end there. R realizations from a random variable X raised to the Nth power is expressed symbolically as  $\{X_1 * X_1 * X_1 * \dots_{N \text{ times}} \dots, X_2 * X_2 * X_2 * \dots_{N \text{ times}} \dots, \dots, X_R * X_R * X_R * \dots_{N \text{ times}} \dots\}$ , which is quite different from how the Lognormal is being concocted (the end of Chapter 76 gives a detailed expression of the Lognormal). Much lower variability is noted here within each element as compared with the Lognormal, and log density of this power process is not Normal at all. Rather, it rises steadily on the left upward throughout its range, only to fall off suddenly and sharply towards the end on the right.

The Normal distribution is encountered in repeated additions of any random variable.  $\text{Normal} = X_1 + X_2 + X_3 + \dots + X_N$  where  $X_1$  are independent and identical realizations of X defined anywhere on  $(-\infty, +\infty)$ , which then calls into play the Central Limit Theorem guaranteeing the Normality of the sum in the limit as N gets large, regardless of distribution form or parameter of X.

Alternatively, the Lognormal distribution is a random variable defined over  $(0, +\infty)$  and whose logarithm (of any base) is the Normal distribution. Applying base e, its expression becomes:  $\text{Lognormal} = e^{\text{Normal}} = e^{(X_1+X_2+X_3+\dots+X_N)} = e^{(X_1)} * e^{(X_2)} * e^{(X_3)} * \dots * e^{(X_N)}$ . This latter definition then implies that the Lognormal distribution can be represented as a process of repeated multiplications, as was mentioned earlier. In brief:

$$\text{Lognormal}(\text{location, shape}) = e^{\text{Normal}(\text{location, shape})}$$

$$\text{Normal}(\text{mean, s.d.}) = \ln(\text{Lognormal}(\text{mean, s.d.}))$$

The shape and location parameters of the Lognormal are the standard deviation and the mean, respectively, of the ‘generating’ Normal distribution. Simulations

of the Lognormal distribution for the purpose of examining its digital behavior give reasonable logarithmic behavior whenever the shape parameter is roughly over 0.4, regardless of the value assigned to the location parameter. Better results are obtained for higher values of the shape parameter. A near-perfect logarithmic behavior is observed whenever shape parameter is approximately over 1.0, and this result is totally independent of the value assigned to the location parameter. For shape parameter in the range between 0 and 0.4 approximately, there are strong fluctuations in its digital distribution that also depends to a large extent on the location parameter, and for the most part doesn't resemble the logarithmic at all. It is worth noting that for very low values of the shape parameter the Lognormal distribution is almost symmetrical, and it becomes progressively more skewed to the right as the shape parameter increases in value (hence the name 'shape'), coinciding with progressively better conformity to the logarithmic. This is nicely consistent with what was noted earlier, namely that an asymmetrical one-sided tail to the right is an essential feature in all logarithmic data sets and distributions. This is also consistent with Related Log Conjecture, which requires a wide range on related log-axis, a range that is directly proportional here to the shape parameter being the standard deviation of the 'generating' Normal distribution (higher value of s.d. for the Normal implies wider range there). Since translations of related log curve with wide range to the right or to the left do not affect logarithmic behavior, the value of the location parameter (the mean of the 'generating' Normal distribution) is immaterial. Low values of the shape parameter represent situations where 'generating' Normal is focused too narrowly on some particular small range, implying narrow range on related log-axis and low order of magnitude for the Lognormal. Employing a different perspective, one may summarize all the above as follows: the Normal distribution being a model of a variable obtained from numerous additions of independent and identical factors is not logarithmic; the Lognormal, being a model of a variable obtained from numerous multiplications of independent and identical factors, is indeed logarithmic.

Finally, while  $k/x$  distribution is exactly logarithmic when defined over proper range, the Lognormal is only very nearly perfectly logarithmic for high values of the shape parameter, and this fact is surely more than sufficient for all practical purposes for the data analyst, fraud detector, and the practical statistician. Yet theoretically the Lognormal can never attain an exact logarithmic behavior in a purely mathematical sense if one is being too strict in measuring it [microscopic infinitesimal 'deviations' are observed by experienced pure mathematicians with

excellent vision]. The Lognormal is logarithmic only in a limiting sense when the shape parameter approaches infinity. In the same vein, Related Log Conjecture can never yield an exact logarithmic behavior in the eye of the strict and stern pure mathematician, no matter how wide the range on the log-axis may be, no matter how smooth and gradual the curve appears, only in a limiting sense.

## SCRUTINIZING DIGITS WITHIN LOGNORMAL, EXPONENTIAL, AND K/X

---

Let us examine more closely digital behavior of the Lognormal distribution, not as traditionally done for the whole range combined (**globally**), but rather (**locally**) in separate stages for all relevant sub-intervals bordered by adjacent integral powers of ten, carefully scrutinizing those mini digital distributions in order to be able to listen to their own individual digital messages.

Figure 4.63 shows mini digit distributions between adjacent IPOT, interval by interval, for simulations of the Lognormal with shape parameter 1.11 and location parameter 2.303. The simulation generated a total of 13,000 values or realizations (called cases) for the entire curve. Only 257 values (cases) fell inside the left-most sub-interval  $[0.1, 1]$ , representing merely 257/13,000 or 2.0% of overall data (i.e. weight of sub-interval). Digital configuration of those 257 values falling inside  $[0.1, 1]$  considered in isolation showed a strong tendency to favor high digits. Such an inverse configuration where only two cases start with digit 1 [2/257 or 0.8%] and 47 cases start with digit 9 [47/257 or 18.3%] is in sharp contrast to the logarithmic configuration of the entire Lognormal data set of 13,000 cases (shown on the last column on the rightmost side of the table.)

The pattern seen in Fig. 4.63 is consistent across all Lognormal distributions having other parameter sets compatible with digital logarithmic behavior (i.e. high shape), so all this is quite general. On intervals near the origin, low digits lose leadership, and curve is actually ascending. Yet this surprising strength of high digits is fleeting and only over a very small portion of overall data. Soon afterwards, digits rapidly achieve equality, followed by a strong lead of low digits where the logarithmic is achieved locally. Finally, a near dominance of low ones over high ones is observed at the end on the far right. Miraculously, overall leading digits distribution of the Lognormal in the aggregate is almost perfectly logarithmic in spite of that (digital) emotional roller coaster.

Let us scrutinize mini digit distributions for the exponential distribution as well. The exponential distribution with parameter  $p > 0$  is defined as  $f(x) = pe^{-px}$

Digit	[0.1, 1]	[1, 10]	[10, 100]	[100, 1000]	OVERALL
	=====	=====	=====	=====	=====
1	0.8	11.1	49.0	78.8	30.5
2	2.3	13.4	21.3	14.1	17.0
3	5.1	13.4	11.1	3.9	11.9
4	9.3	13.0	7.2	2.4	9.9
5	10.1	12.1	4.4	0.4	8.1
6	16.7	11.4	3.0	0.0	7.2
7	18.3	10.0	2.0	0.4	6.1
8	19.1	8.2	1.1	0.0	4.8
9	18.3	7.4	1.0	0.0	4.4
<b>Cases:</b>	257	6,222	6,266	255	13,000
<b>Interval's Weight:</b>	2.0%	47.9%	48.2%	2.0%	100%

Figure 4.63 Mini Digital Configurations of the Lognormal on IPOT Sub-Intervals

Digit	[0.01, 0.1]	[0.1, 1]	[1, 10]	[10, 100]	OVERALL
	=====	=====	=====	=====	=====
1	11.6	11.2	14.4	50.6	31.8
2	9.9	11.7	13.3	24.8	18.9
3	14.3	11.3	12.3	12.3	12.7
4	12.3	11.3	11.8	6.3	9.1
5	8.9	10.7	11.1	3.0	7.1
6	9.1	11.4	10.2	1.6	6.0
7	12.3	10.2	9.4	0.8	5.1
8	10.6	11.0	9.1	0.3	4.6
9	11.1	11.1	8.4	0.2	4.6
<b>Cases:</b>	88	678	5,865	6,333	13,000
<b>Interval's Weight:</b>	0.7%	5.2%	45.1%	48.7%	100%

Figure 4.64 Mini Digital Configurations of the Exponential on IPOT Sub-Intervals

over the positive range  $(0, +\infty)$ . It has an asymmetrical one-sided long tail to the right that constantly falls off throughout its entire range. Digit distribution of the exponential is not exactly logarithmic but rather close to it. No matter what value parameter  $p$  takes, overall digit distribution considered over its entire range of  $(0, +\infty)$  is quite close to the logarithmic.

The table in Fig. 4.64 shows mini digit distributions between adjacent IPOT, interval by interval, for 13,000 simulated realizations from the exponential with parameter 0.069315. Thirty seven numbers falling below 0.01 and above 100 were discarded.

Except for the narrow sub-interval near the origin where data is quite thin and unreliable, the pattern seen in Fig. 4.64 is totally consistent across all exponential distributions having other parameters, so all this is quite general. Unlike the Lognormal, here low digits always win over high digits on each of these sub-intervals, given that plenty of simulated values are generated. Yet their advantage over high digits is weak near the origin and gets progressively stronger later on, culminating finally at the end towards the far right in extreme dominance over high digits. Since the slope of the exponential is negative everywhere as seen from its derivative  $f'(x) = -p^2 e^{-px}$ , therefore the curve is always falling, with the obvious result that high digits can never hope to win the leadership game even just locally somewhere. Moreover, even though slope is constantly getting flatter (since second derivative is  $+p^3 e^{-px}$ ), nonetheless low digits are continuously taking more leadership from high digits as we move right and IPOT sub-intervals become much wider.

A very different pattern altogether is found in the case of  $k/x$  distribution. It was shown earlier that  $k/x$  distribution exhibits the logarithmic property steadily and consistently throughout its entire range; hence perhaps a perfect repetition of the logarithmic configuration should be empirically found in all sub-intervals and segments of the range. The table in Fig. 4.65 shows mini digit distributions between adjacent IPOT, interval by interval, for 13,000 simulations from  $0.10857/x$  distribution defined on (1, 10000). In stark contrast to the Lognormal and the exponential,  $k/x$  distribution has indeed consistent and steady logarithmic digital configuration throughout. In addition, exactly equal portions of overall data fall on those IPOT sub-intervals, 25% for each of the four segments in this simulation example.

It should be noted that in Figs. 4.63, 4.64, and 4.65, overall digit distribution (the last column on the right) cannot be calculated simply as the straight average of all the mini digit distributions on IPOT sub-intervals. Rather, it can be calculated as the weighted average, weighted by the percent/portion of data falling within each IPOT sub-interval. On another note, the potential overlapping which may occur at the edges for 0.1, 1, 10, 100, 1000 are ignored and in any case no data points have fallen exactly on them here. The standard practice in all such digital examination is to have all left edges open and all right edges closed, as in  $[L, R)$ , so as not to overlap any points, as in  $[1, 10)$ ,  $[10, 100)$ ,  $[100, 1000)$ .

Digit	[1, 10]	[10, 100]	[100, 1000]	[1000, 10000]	OVERALL
1	30.1	30.1	30.1	30.1	30.1
2	17.6	17.6	17.6	17.6	17.6
3	12.5	12.5	12.5	12.5	12.5
4	9.7	9.7	9.7	9.7	9.7
5	7.9	7.9	7.9	7.9	7.9
6	6.7	6.7	6.7	6.7	6.7
7	5.8	5.8	5.8	5.8	5.8
8	5.1	5.1	5.1	5.1	5.1
9	4.6	4.6	4.6	4.6	4.6
<b>Cases:</b>	3,250	3,250	3,250	3,250	13,000
<b>Interval's Weight:</b>	25%	25%	25%	25%	100%

Figure 4.65 Mini Digital Configurations of  $k/x$  on IPOT Sub-Intervals

Let us examine the interaction between density curves and digital configurations in total generality for any continuous probability distribution. For each sub-interval bordered by two adjacent IPOT points such as (1, 10) or (10, 100), complete uniformity of digital distribution (of first and all higher orders) locally, i.e. non-logarithmic digital equality, implies uniformity of the density function itself (a flat curve), and vice versa, assuming smoothness and continuity. A logarithmic digital distribution on such sub-interval implies a tail to the right with density falling off just as steeply as in the  $k/x$  case. And cases where digit 1 leads much more than its rightful proportion of 30.1%, and where high digits are almost totally excluded from leadership, are areas where the density falls even more rapidly or sharply than  $k/x$ . Clearly, overall digit distribution is obtained after averaging out (the weighted) different shapes/slopes along the entire curve, incorporating all such sub-intervals. It should be emphasized, though, that our old familiar concept of slope from calculus does not translate here literally in the digital context. It is certainly true though that for any two competing density curves of equal average height on any given interval between adjacent IPOT points, the one with steeper negative slope endows more leadership to low digits. Yet slope can't be totally divorced and independent from interval's location on the x-axis or from its height as far as its effects on digital configuration is concerned. A given negative slope value (calculus-wise) over an adjacent IPOT interval has by far weaker and diluted effect on digital configuration if curve is hanging very high far from the x-axis,

while the same value of slope has an enormous digital effect if curve is low near the x-axis, giving by far more advantage to low digits. Resultant digital configuration is derived from the triple interactions of (I) slope, (II) height above the x-axis, (III) location on the x-axis. A good conceptual example about the interplay between these three factors is  $k/x$  distribution, where derivative is  $-k/x^2$ , hence slope is constantly negative but getting flatter even though local mini digit distributions on those sub-intervals between adjacent IPOT points remains constant throughout.

For a visual understanding of the distinct manners in which the above-mentioned three distributions are digitally and graphically configured, the chart in Fig. 4.66 depicts all three of them superimposed. The relationship between density slope and digital behavior shall be clearly demonstrated via these three curves. The focus in all three distributions in Fig. 4.66 is on two specific IPOT sub-intervals, namely (1, 10) and (10, 100). For visual clarity only the interval (0, 70) is shown and luckily very little else is missing by neglecting to show the portion over 70. The choice of parameters is deliberate so as to have all three distributions with the same median value of 10, which enables us to obtain meaningful comparisons

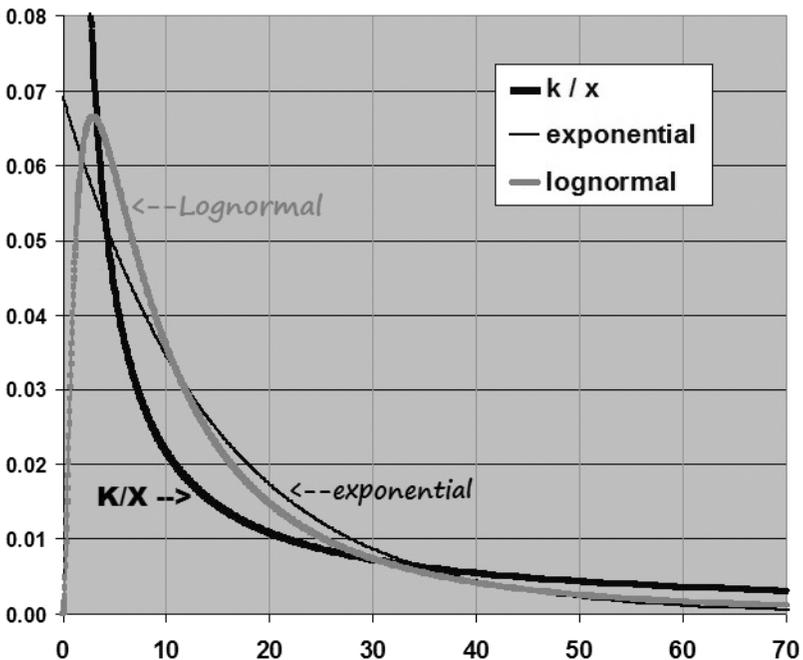


Figure 4.66 The Exponential, Lognormal, and  $k/x$  Distributions Centered on 10

between them. A slightly different  $k/x$  distribution will be considered here so as to bring the median around to 10, therefore for Fig. 4.66,  $k/x$  is defined over  $(1, 100)$ , namely  $(10^0, 10^2)$ , so that  $k = 1/[G*\ln 10] = 1/[2*\ln 10] = 0.21715$ . Median in general for  $k/x$  defined over  $(10^S, 10^{S+G})$  is  $10^{[G/2 + S]}$ , hence the median here is  $10^{[2/2 + 0]}$ , namely 10. The Lognormal distribution over  $(0, +\infty)$  is the same as of Fig. 6.63 with 2.303 location parameter and 1.11 shape parameter. The median for the Lognormal is always  $e^{\text{location-parameter}}$ , namely  $e^{2.303} = 10$  in our case. The exponential distribution over  $(0, +\infty)$  here is the same as of Fig. 4.64 with parameter 0.069315. The median for the exponential is always  $(1/p)*\ln(2)$ , hence median here is  $(1/0.069315)*\ln(2)$  or simply 10. Aggregate digit distributions of the exponential and the Lognormal distributions over the limited segment of  $(1, 100)$  are close to being logarithmic and this can be easily confirmed via Figs. 4.63 and 4.64 by simply averaging out these two columns both having about the same data weight. Digit configuration of  $k/x$  distribution is perfectly logarithmic over  $(1, 100)$ . To recap: all three curves on  $(1, 100)$  are just about logarithmic, albeit arriving at it in three distinct ways as shall be shown in the next paragraph. [Note: The chart in Fig. 4.66 terminates abruptly at 70, neglecting to show only 8% of the area for the rest of  $k/x$  distribution to the right, just 4% for the rest of the Lognormal, and a mere 1% for the rest of the exponential, so that the overall picture can still be visualized clearly. Also, the abrupt termination of the roof at 0.08 masks just a slight rise of  $k/x$  up to 0.21 when  $x$  is 1].

All three distributions need to transit these two sub-intervals between adjacent IPOT points — namely  $(1, 10)$  &  $(10, 100)$  — in a way that would result in the logarithmic or nearly so overall. While  $k/x$  has a steady fall (digit-wise) over these two sub-intervals, the exponential starts out around  $(1, 10)$  not as steep as  $k/x$  (hence digits are more equal there), only to reverse course later on  $(10, 100)$  and to descend even more intensely than  $k/x$  does (hence digits are more skewed there). The Lognormal which is actually ascending briefly around  $(1, 3)$ , letting high digits enjoy a momentary and delusional victory, quickly reverses course and bows to the inevitable fall, then finally joins the exponential around  $(10, 100)$  in its precipitous descent steeper than that of  $k/x$  distribution resulting in the supreme dominance of low digits there. This overall description of the three curves in Fig. 4.66 perfectly corresponds to the three digital configurations interval-by-interval shown in Figs. 4.63, 4.64, and 4.65, and both vistas (slopes and digits) turned out to be nicely consistent with each other, as they should.

## LEADING DIGITS INFLECTION POINT

---



---

Let us examine related log for each of the three distributions in the previous chapter. The common (decimal) log density of  $k/x$  distribution is Uniform, as seen in Proposition I. The common log density of the Lognormal is Normal as in its definition, regardless of the base employed, be it 10,  $e$ , or any other base. Schematically, related common log of the Lognormal is  $\text{LOG}_{10}(e^{\text{Normal}})$ , a mix of natural and common logs and bases. Using the log identity  $\text{LOG}_A(X) = \text{LOG}_B(X)/\text{LOG}_B(A)$  twice, we get:  $\text{LOG}_e(e^{\text{Normal}})/\text{LOG}_e(10) = \text{Normal}/\text{LOG}_e(10) = \text{Normal}/[\text{LOG}_{10}(10)/\text{LOG}_{10}(e)] = (\text{LOG}_{10}e)*\text{Normal}$ , which is simply another Normal distribution representing a multiplicative transformation of the original Normal(location, shape). The common log density of the exponential with parameter  $p$  is distributed as  $f(y) = p*\ln 10*[10^y]*[e^{-p*(10 \text{ to the } y)}]$ . This is derived by employing the Distribution Function Technique for transformations. As can be deduced from the latter algebraic expression, all related log densities of the exponential and regardless of parameter  $p$  value are asymmetrical. They appear as a contorted bell-shape-like curve skewed to the left and its entire range is always approximately 2.5 units on the log-axis. As conjectured earlier with regard to Related Log Conjecture, the system can't deliver strong logarithmic behavior for asymmetrical related log densities unless a bit of a wider spread on the log-axis is available. Since all exponential distributions suffer from their own log-asymmetry, and since this deficiency is never corrected by any additional log-axis spread, their digital configurations are slightly different from the logarithmic. Figure 4.67 depicts related (common) log curves for the three distributions shown in Fig. 4.66. Since all three distributions themselves were centered on 10, their related log densities are centered on  $\log(10)$ , namely centered on 1 as can be clearly seen in Fig. 4.67.

In general for any data or distribution, the point on related (common) log curve having maximum height is called **Leading Digit Inflection Point (LDIP)**, corresponding to an important turning point in digital behavior. Assuming there are no dents, hills, and valleys in the log curve, or that second

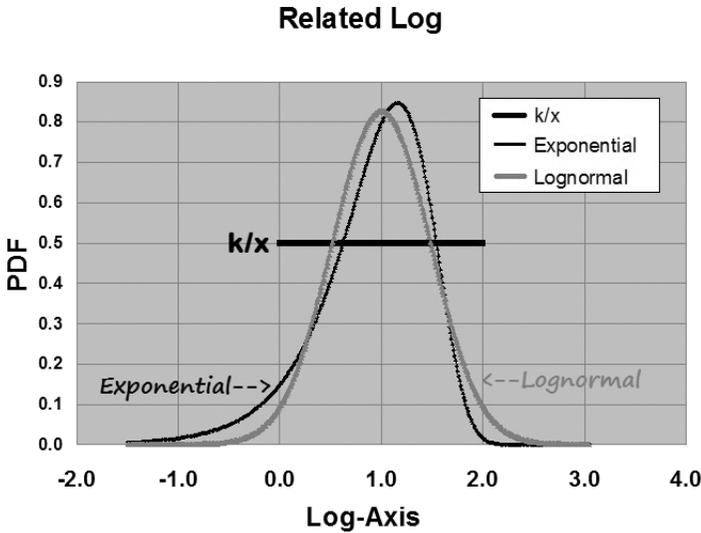


Figure 4.67 Related Log of the Exponential, Lognormal, and  $k/x$  Centered on 10

derivative is continuously negative with slope steadily decreasing as in the case of the semi-circle for example, then LDIP is unique. That infinitesimal max point of related log curve corresponds to a point of the distribution itself of perfect logarithmic condition, borrowing momentarily from  $k/x$  distribution its logarithmic property (Proposition II). Exactly at the max, log curve is temporarily flat and horizontal, having zero slope and thus suggestive of course of the uniform distribution if only at that infinitesimal location. To the left of LDIP, high digits are faring better than the typical logarithmic condition, overcoming its bias, and may achieve equality or even dominance. To the right of LDIP, dichotomy between high and low digits is even more acute than that of the logarithmic inequality. LDIP refers to two manifestations, one on related log curve, and the other on the distribution curve itself; both are mirror images of each other. Note that the largest weight or portion of overall data is typically given to that IPOT sub-interval containing LDIP.

To find LDIP of the exponential, we set the derivative of its related log function to zero and solve for  $y$ .

$$\begin{aligned}
 f'(y) &= (p \cdot \ln 10 \cdot (10^y) \cdot (e^{-p \cdot (10 \text{ to the } y)}))' = 0 \\
 p \cdot \ln 10 \cdot [10^y \cdot (e^{-p \cdot (10 \text{ to the } y)}) \cdot (-p \cdot \ln 10 \cdot 10^y) + (e^{-p \cdot (10 \text{ to the } y)}) \cdot \ln 10 \cdot 10^y] &= 0 \\
 p \cdot \ln 10 \cdot (10^y) \cdot (e^{-p \cdot (10 \text{ to the } y)}) \cdot \ln 10 \cdot [(-p \cdot 10^y) + 1] &= 0
 \end{aligned}$$

$$\begin{aligned} [(-p*10^y) + 1] &= 0 \\ 1 &= p*10^y \\ 1/p &= 10^y \end{aligned}$$

Hence  $\text{LOG}_{10}(1/p)$  is LDIP, the point on the log-axis yielding maximum height for the log density. This point corresponds to  $1/p$  on the x-axis of the exponential density curve itself.

In Fig. 4.67 for the exponential with parameter 0.069315, this point should be  $\text{LOG}_{10}(1/0.069315)$  or 1.16, which agrees visually. For the exponential itself,  $1/p$  is  $1/(0.069315)$  or 14.4, which is nicely compatible with Figs. 4.66 and 4.64. The table in Fig. 4.64 clearly tells of a more equal digital configuration to the left of 10 (which is roughly near 14.4 inflection point), and more extreme skewness to the right of 10.

To find LDIP of the Lognormal, we maximize related (common) log of its function, namely  $\text{Max}\{\text{density of } \text{LOG}_{10} \text{ Lognormal}\} =$

$$\text{Max}\{\text{density of } \text{LOG}_{10}(e^{\text{Normal}})\} =$$

$\text{Max}\{\text{density of } (\text{LOG}_{10} e) * \text{Normal}\}$ . This maximization is achieved whenever the generating Normal itself is at max, namely when it's resting at its mean parameter (since the mode being the highest point or max, the median, and the average are all the same for any Normal distribution). Therefore, the point on the log-axis of related log of the Lognormal yielding maximum height is simply: LDIP =  $(\text{LOG}_{10} e) * (\text{location-parameter})$ . Finally, on the curve of the Lognormal itself, Leading Digits Inflection Point is at:

$$\begin{aligned} &10 \text{ to the power of } [(\text{LOG}_{10} e) * (\text{location parameter})] \\ &\text{Utilizing the identity } B^{(R * Q)} = (B^R)^Q \text{ we get:} \\ &[10 \text{ to the power of } \text{LOG}_{10} e]^{\text{location parameter}} = \\ &[e]^{\text{location parameter}} \end{aligned}$$

In Fig. 4.67 for the Lognormal with 2.303 as the location parameter, LDIP on the log-axis should be  $\text{LOG}_{10}(2.718282) * 2.303$  or 1.0002, which agrees visually. For the Lognormal itself,  $e^{\text{location parameter}}$  is  $e^{2.303}$  or 10.004 which is nicely compatible with Figs. 4.63 and 4.66. The table in Fig. 4.63 clearly tells of a more equal digital configuration to the left of 10, and extreme skewness even more than the logarithmic condition to the right of 10.

For  $k/x$  distribution, on the other hand, LDIP does not exist;  $k/x$  is continuously and consistently logarithmic, and its related log density is flat and uniform throughout.

The distinction made in the last two chapters regarding the manner in how these three distributions arrive at the logarithmic is not merely conceptual, but also quite applicable to pieces cut out from the whole of distributions. For the Lognormal and the exponential, all the parts are needed from the whole in order to obtain that aggregate (near) logarithmic behavior over its entire range. It is not permitted to cut out and focus on any of its sub-intervals and expect logarithmic behavior. The  $k/x$  distribution stands out conspicuously as the only one we are permitted to cut pieces out and still observe logarithmic behavior — so long as these sub-intervals come with an integral exponent difference.

## DIGITAL DEVELOPMENT PATTERN FOUND IN ALL REAL-LIFE RANDOM DATA

---

---

We now have the tools to gain better insight into how typical real-life random data sets behave digitally in general along the path of their entire ranges, be they logarithmic or otherwise. This is done by integrating a variety of essential results seen earlier, namely the conjecture that Hill's super distribution relates to or resembles the Lognormal; that related log densities of single-issue physical data, Random Linear Combinations, and chains of distributions, all resemble the Normal or the contorted semi-circular; the dichotomy between the deterministic and the random; and the closer scrutiny given to the Lognormal, exponential, and  $k/x$  distributions regarding their digital behavior on smaller IPOT sub-intervals within their entire range. That closer examination in the case of the Lognormal showed that digital behavior on the left of inflection point is quite different than digital behavior on the right of that point, therefore it is natural to conjecture that the same digital variation and development should be found in general for all random real-life data. Indeed, empirical examinations of real-life random data of all types consistently confirm this conjecture. The examinations of numerous data sets such as accounting data, U.S. census data, scientific data, RLC, chains, abstract and real simulations of Hill's super distribution, and so forth, all show that same digital development pattern! There is not a single exception! On the other hand, computer simulations of exponential growth series (which relate to  $k/x$  via Proposition VI) show no digital development pattern whatsoever, but rather a steady and constant logarithmic behavior across all IPOT sub-intervals.

The 2009 census data seen earlier on population of 19,509 incorporated cities and towns in the USA shall be examined for any possible digital development. The table in Fig. 4.68 demonstrates mini digit distributions on all relevant IPOT sub-intervals, except for the last one on the extreme right of (1000000, 10000000) which was omitted due to very low proportion count, namely just nine data points of mega cities representing 0.05% of overall data. Along the sub-intervals, as we move the focus to higher values from the left regions towards the right regions,

<b>Left Border Point</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>	<b>100,000</b>
<b>Right Border Point</b>	<b>10</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>	<b>100,000</b>	<b>1,000,000</b>
	===	===	====	=====	=====	=====
<b>Digit 1</b>	<b>14.8</b>	<b>5.3</b>	<b>19.1</b>	<b>37.3</b>	<b>46.0</b>	<b>62.9</b>
<b>Digit 2</b>	<b>7.4</b>	<b>8.1</b>	<b>17.4</b>	<b>19.7</b>	<b>20.2</b>	<b>17.6</b>
<b>Digit 3</b>	<b>3.7</b>	<b>7.0</b>	<b>13.6</b>	<b>11.6</b>	<b>10.9</b>	<b>6.0</b>
<b>Digit 4</b>	<b>7.4</b>	<b>9.2</b>	<b>11.5</b>	<b>8.6</b>	<b>6.4</b>	<b>4.1</b>
<b>Digit 5</b>	<b>7.4</b>	<b>11.5</b>	<b>9.9</b>	<b>6.3</b>	<b>5.8</b>	<b>3.0</b>
<b>Digit 6</b>	<b>14.8</b>	<b>13.9</b>	<b>8.8</b>	<b>5.3</b>	<b>3.8</b>	<b>3.0</b>
<b>Digit 7</b>	<b>7.4</b>	<b>13.9</b>	<b>7.6</b>	<b>4.3</b>	<b>2.8</b>	<b>1.5</b>
<b>Digit 8</b>	<b>14.8</b>	<b>17.0</b>	<b>6.1</b>	<b>4.0</b>	<b>2.3</b>	<b>1.1</b>
<b>Digit 9</b>	<b>22.2</b>	<b>14.1</b>	<b>6.0</b>	<b>2.9</b>	<b>1.7</b>	<b>0.7</b>
	-----	-----	-----	-----	-----	-----
<b>Data points:</b>	<b>27</b>	<b>1065</b>	<b>8202</b>	<b>7285</b>	<b>2654</b>	<b>267</b>
<b>% Overall Data</b>	<b>0.14%</b>	<b>5.5%</b>	<b>42.0%</b>	<b>37.3%</b>	<b>13.6%</b>	<b>1.4%</b>

**Figure 4.68** Consistent Digital Development Pattern Seen in U.S. Population Data

low digits are continuously and steadily obtaining more leadership from high digits. This steady evolution of digital proportions along the x-axis is what characterizes ‘random’ and statistical processes, differentiating it from the ‘deterministic’, and this is clearly demonstrated here in this population data. [Note: defined sub-ranges overlap at 10, 100, 1000 etc., but since only very few values in the data set are exactly IPOT the issue is not significant. In any case, the standard rule for IPOT values is to place them always on the right sub-interval].

Let us visually demonstrate this digital evolution directly as it occurs in the raw values of the population data itself. For that purpose, population’s histogram as well as the generic type of Benford  $k/x$  density between IPOT points shall be superimposed on the same scatter plot, so that the contrast between these two different digital conditions (pure Benford type and population data) could be clearly observed on the most primitive level to dispel all doubts. Three IPOT sub-intervals would be considered for this comparison, namely (10, 100), (100, 1000), and (1000, 10000), upon which three separate mini histograms are to be constructed. Since it was conveniently (arbitrarily) decided to construct all three mini histograms uniformly with exactly 30 bins (30 rectangles) each, therefore each histogram is based on 3, 30, and 300 units of rectangular width (bin width), respectively.

In Fig. 4.69, the rising histogram of U.S. population data on the sub-interval (10, 100) is sharply contrasted with the falling  $k/x$  curve. This is perfectly consistent with the reversal of the local mini leading digits configuration shown in Fig. 4.68 in the second column of (10, 100) favoring high digits.

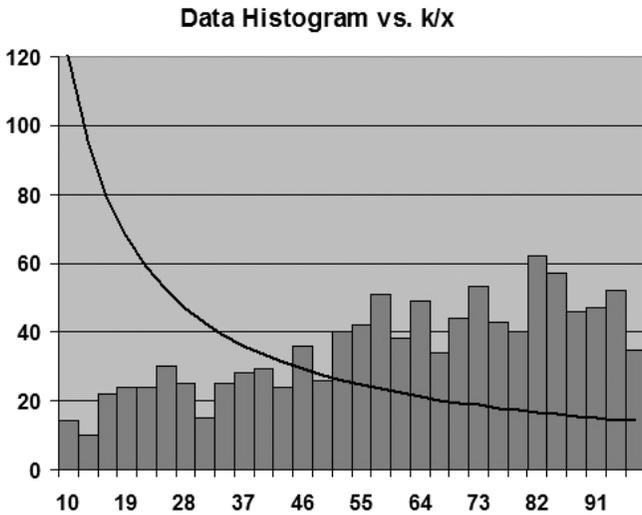


Figure 4.69 Rising Histogram on (10, 100) Consistent with Digital Reversal

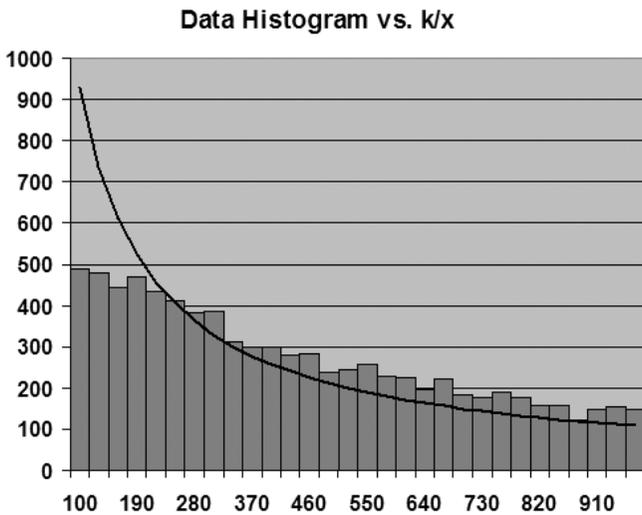
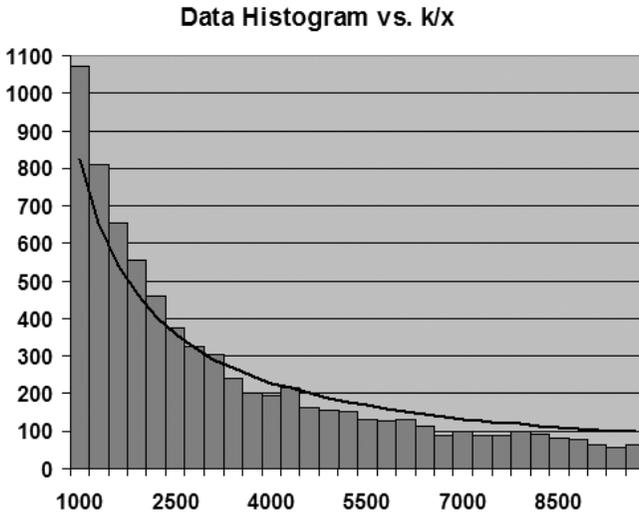


Figure 4.70 Mild Fall on (100, 1000) Consistent with Moderate Digital Skewness

In Fig. 4.70, the gently falling histogram of U.S. population data on the sub-interval (100, 1000) is contrasted with the much steeper fall of the  $k/x$  curve. This is perfectly consistent with the local leading digits configuration shown in Fig. 4.68 in the third column of (100, 1000) which mildly favors low digits over high ones, and where digits are less skewed compared with the logarithmic configuration.



**Figure 4.71** Sharp Fall on (1000, 10000) Consistent with Super Skewed Digits

In Fig. 4.71, the precipitously falling histogram of U.S. population data on the sub-interval (1000, 10000) is contrasted with the much gentler and milder fall of the  $k/x$  curve. This is perfectly consistent with the local leading-digits configuration shown in Fig. 4.68 in the fourth column of (1000, 10000) which strongly favors low digits over high ones, over and above the logarithmic configuration.

In order to be able to superimpose a fitting  $k/x$  density curve for each of the three histograms above, the value of  $k$  is chosen deliberately and differently so  $k/x$  is not way below or way above a given histogram. This is done by simply equating area under the  $k/x$  curve to the area of all the rectangles of a given histogram, because it is the *area* in statistical mathematics that represents probability, and probability is what we wish to equate here.

The following construction puts both curves on an equal footing:

$$\int k/x \, dx = \text{total area of the mini histogram of the particular sub-interval}$$

$$\int k/x \, dx = \sum [\text{width of rectangle } i * \text{height of rectangle } i]$$

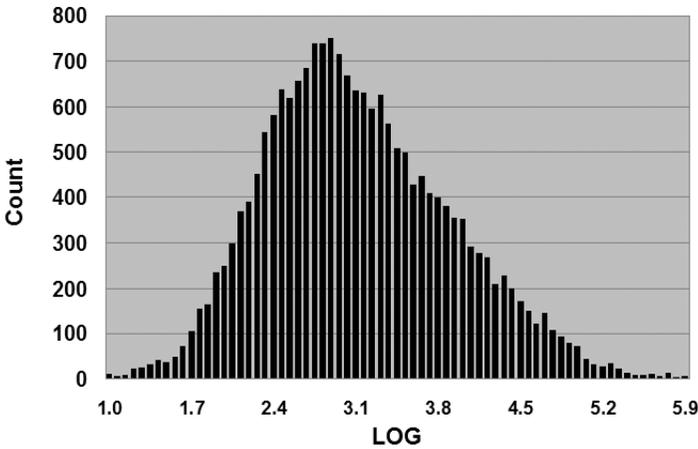
$$k = \sum [\text{width of rectangle } i * \# \text{ of pop values falling within rectangle } i] / \int 1/x \, dx$$

$$k = [\text{rectangle's width}] * (\sum [\# \text{ of pop values within rectangle } i]) / \int 1/x \, dx$$

*since bin width is uniform across all rectangles within a single mini histogram*

$$k = [\text{rectangle's width}] * (\# \text{ of pop values in entire sub-interval}) /$$

$$[\ln(\text{upper}) - \ln(\text{lower})]$$



**Figure 4.72** Related Log Histogram of U.S. Cities and Towns Population Data

For example, for the last chart in Fig. 4.71 of sub-interval (1000, 10000), the histogram is based upon a rectangular width of 300; the total heights of all the rectangles is 7285 (Fig. 4.68, fourth column), hence the value of  $k$  is calculated as:  $k = [300 * 7285] / [\ln(10000) - \ln(1000)] = 949150.6$ .

An examination of the histogram of related log of U.S. population centers data shown in Fig. 4.72 should reveal the basis of its development pattern — in the spirit of Related Log Conjecture. While there is no good resemblance to either the Normal curve or the semi-circular one (only to the asymmetrical triangular perhaps), the histogram starts and ends gradually on the log-axis itself without any abrupt spikes, and it clearly shows two distinct regions: the one on the left rising, and the one on the right falling. Also, the strong logarithmic behavior of the data can be clearly predicted and explained by the shape of the log histogram and the fact that the spread on the log-axis is more than sufficiently wide, with a comfortable 4.0 log-axis units. Inflection point here is around the log value of 2.9, which corresponds to the population value of 794 for the data itself [calculated as  $10^{2.9}$ ]. This value of 794 is consistent with the digital development pattern seen for this data set in Fig. 4.68, as mini digit distributions are less skewed and milder than the Benford condition to the left of this inflection point (1000 approximately) and more extremely skewed to the right of it.

A markedly different situation is seen in all computer simulations of exponential growth series, where no development whatsoever appears, as mini digit distributions are steady and constant everywhere throughout the entire range. The table

<b>Left Border Point</b>	<b>10</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>	<b>100,000</b>	<b>1,000,000</b>
<b>Right Border Point</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>	<b>100,000</b>	<b>1,000,000</b>	<b>10,000,000</b>
	====	=====	=====	=====	=====	=====
<b>Digit 1</b>	<b>30.8</b>	<b>30.8</b>	<b>30.8</b>	<b>30.8</b>	<b>30.8</b>	<b>29.5</b>
<b>Digit 2</b>	<b>17.9</b>	<b>17.9</b>	<b>16.7</b>	<b>16.7</b>	<b>16.7</b>	<b>17.9</b>
<b>Digit 3</b>	<b>11.5</b>	<b>11.5</b>	<b>12.8</b>	<b>12.8</b>	<b>12.8</b>	<b>12.8</b>
<b>Digit 4</b>	<b>10.3</b>	<b>10.3</b>	<b>10.3</b>	<b>10.3</b>	<b>10.3</b>	<b>9.0</b>
<b>Digit 5</b>	<b>7.7</b>	<b>7.7</b>	<b>7.7</b>	<b>7.7</b>	<b>7.7</b>	<b>9.0</b>
<b>Digit 6</b>	<b>6.4</b>	<b>6.4</b>	<b>6.4</b>	<b>6.4</b>	<b>6.4</b>	<b>6.4</b>
<b>Digit 7</b>	<b>6.4</b>	<b>6.4</b>	<b>6.4</b>	<b>6.4</b>	<b>5.1</b>	<b>5.1</b>
<b>Digit 8</b>	<b>5.1</b>	<b>5.1</b>	<b>5.1</b>	<b>5.1</b>	<b>5.1</b>	<b>5.1</b>
<b>Digit 9</b>	<b>3.8</b>	<b>3.8</b>	<b>3.8</b>	<b>3.8</b>	<b>5.1</b>	<b>5.1</b>
	-----	-----	-----	-----	-----	-----
<b>Data points:</b>	<b>78</b>	<b>78</b>	<b>78</b>	<b>78</b>	<b>78</b>	<b>78</b>
<b>% Overall Data</b>	<b>16.7%</b>	<b>16.7%</b>	<b>16.7%</b>	<b>16.7%</b>	<b>16.7%</b>	<b>16.7%</b>

Figure 4.73 Steady Digital Configurations — No Development in Exponential Growth

in Fig. 4.73 depicts digital configurations on each of the relevant IPOT sub-intervals for the first 468 elements of 3% exponential growth series, starting at the initial value of 10. Clearly no digital development exists for this series. When other types of (logarithmic) exponential growth series are simulated, they give the same result, namely the same steady logarithmic configuration throughout the entire range, lending additional support to the assertion that this result is entirely general. When the intimate relationship between exponential growth series and  $k/x$  distribution is acknowledged as per Proposition VI, this last result should not surprise us in the least, since steady digital configuration and a complete lack of development was theoretically proven as well as empirically tested for  $k/x$  earlier (as seen in Fig. 4.65). Figure 4.73 only confirms the consistency of the overall understanding developed here of the leading-digit phenomena in general. An examination of the flat and uniform histogram of related log of this exponential 3% growth series (shown in Fig. 4.74) clearly explains its lack of digital development, as well as its nearly perfect logarithmic behavior. The first term in the exponential growth series is 10, and the last 468th term is 9885165.3, hence its related log spans (1.0000, 6.9957), namely nearly an exact integral range of six units on the log-axis, which is also wide enough. This in turn implies a near-perfect logarithmic behavior, as shown in its first leading digits configuration of {30.6, 17.3, 12.4, 10.0, 7.9, 6.4, 6.0, 5.1, 4.3} for the entire series, with an extremely low 0.6 SSD value demonstrating its strong compliance with the law of Benford. [Note: due to the discrete nature of exponential growth series, a large

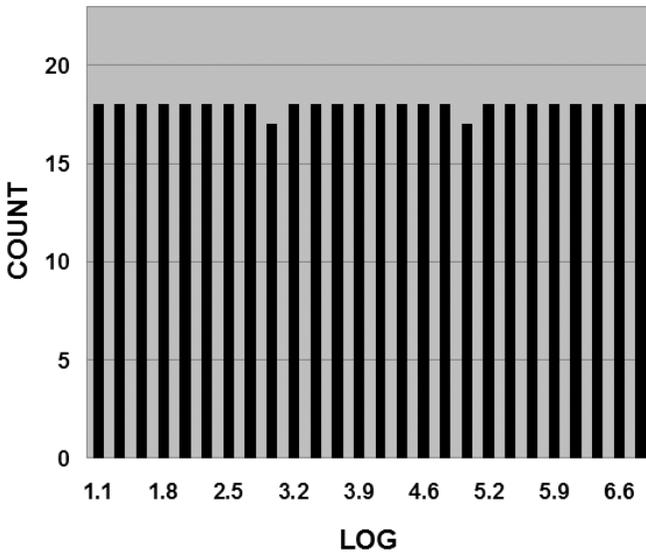


Figure 4.74 Flat Related Log Histogram of Exponential 3% Growth from Base 10

range over log-axis is useful and at times necessary for logarithmic behavior, in addition to the approximate integral span requirement. In contrast, its continuous-mirror-image, namely  $k/x$ , only requires an integral range, not any particularly long span, as encountered earlier in the example of  $k/x$  defined over  $(10, 100)$  with a meager log-axis range of 1 unit].

## DIGITAL DEVELOPMENT PATTERN SEEN ONLY UNDER IPOT PARTITION

---

---

In order to observe digital development pattern in random data, it is crucial to thoughtfully and deliberately decide upon the exact nature of the partitioning of the entire range of data into smaller and adjacent sub-intervals. Otherwise, the patterns cannot be observed.

Clearly, sub-intervals need to be compatible and equal regarding their digital possibilities in order to perform fair comparisons. Sub-intervals do not have to be of the same size in the usual distance/length/range sense, but rather in a digital sense. We seek an equality of digital opportunity between the various sub-intervals, not equality of length between them. It is only in order to obtain equal digital opportunity between the digits themselves that we require at least that each of the nine digits should occupy the same distance/range in the usual sense within any given sub-interval. Thus, the requirement of proper comparisons of digital conditions between sub-intervals necessitates constructing them in one very particular and unique fashion by letting them stand between adjacent integral powers of ten such as  $[1, 10)$ ,  $[10, 100)$ ,  $[10^2, 10^3)$ , and so forth. Sub-intervals constructed in any other way wouldn't do! The improper short sub-interval  $[10, 90)$  does not contain any numbers with first digit 9 no matter what density curve or histogram is hanging above, hence it's too short in digital context and unfairly biased against digit 9. The same difficulty exists with the improper short sub-interval  $[20, 100)$  where digit 1 doesn't have a chance to occur. In contrast, the improper sub-interval  $[10, 200)$  is too long in digital context because here digit 1 gets some undeserved extra advantage, dominating portions  $[10, 20)$  as well as  $[100, 200)$ , leaving by far less ground for digits 2 through 9 irrespective of the nature of the density curve hanging above. On the other hand, the digitally-proper sub-interval  $[10, 100)$ , for example, has that digital-wholeness property that we seek in order to make valid comparisons, and here each digit is given an opportunity to express its leadership equally, being that each dominates an equal portion of the sub-interval. For the sub-interval  $[10, 100)$ , digit 1 dominates  $[10, 20)$ , digit 2

dominates [20, 30), and so forth, and thus each digit is granted equal length and opportunity. Note that with a proper partition along IPOT points such as [1, 10), [10, 100), [100, 1000), for example, [1, 10) comes with the small range of 9, and [10, 100) comes with a range of 90, while [100, 1000) comes with the large range of 900. Yet, this variety of ranges does not adversely affect digital comparison in the least as each digit equally and fairly competes among its peers within each of the three sub-intervals above. In summarizing IPOT partition, it is noted that there is no equality of length between the various sub-intervals, but rather equality of length for all nine digits within any given sub-interval.

Different considerations altogether also preclude choosing sub-intervals between two non-adjacent integral powers of ten such as [10, 1000) — expressed as  $[10^1, 10^3)$  — since such partition obscures pure comparisons. Such misguided partition could potentially confuse the digital narrative occurring on [10, 100) with a totally different one occurring on [100, 1000), therefore these two sections should not be mixed. Aggregating digital accounts here, instead of beneficially summarizing things, would simply mask and obscure the more detailed forensic occurrences happening on each of the two separate sections.

On the face of it, it might be tempting to believe that a more flexible and liberal requirement for an equitable partitioning of the range could be gotten by merely ensuring that an exact unity exponent difference exist at the edges, such as in  $[10^{3.58}, 10^{4.58})$ , [10.93, 109.3), [2.5, 25), [20, 200), and so forth. A decimal shift once to the right increases the number of the left edge by a factor of 10, yielding the right edge. Supposedly, in digital context, all nine digits are given a chance to lead in such cases, since each digit has got a hold on some territory. For example on [20, 200), digit 1 dominates [100, 200), digit 2 dominates [20, 30), digit 3 dominates [30, 40), and so forth. Yet since digit 1 here dominates a territory which is ten times wider than the territory of any other digit, this implies a bias in favor of digit 1 and should be avoided. In contrast, the proper IPOT partitioning fairly allocates equal territory to each of the nine digits within each of the sub-intervals.

**In conclusion, IPOT partition constitutes a very natural choice in digital comparisons, and it is only through this digital lens that digital development can be seen.** Indeed, when other misguided partitions (even those having a unity exponent difference) are experimented with and performed on real-life data sets, the results do not show any discernable pattern, yielding instead some incoherent, confused, or even inverse digital development picture.

<b>Left Border Point</b>	<b>4</b>	<b>40</b>	<b>400</b>	<b>4,000</b>	<b>40,000</b>	<b>400,000</b>
<b>Right Border Point</b>	<b>40</b>	<b>400</b>	<b>4,000</b>	<b>40,000</b>	<b>400,000</b>	<b>4,000,000</b>
	=====	=====	=====	=====	=====	=====
<b>Digit 1</b>	<b>23.6</b>	<b>31.6</b>	<b>29.9</b>	<b>28.2</b>	<b>20.1</b>	<b>11.4</b>
<b>Digit 2</b>	<b>36.3</b>	<b>28.8</b>	<b>15.8</b>	<b>12.4</b>	<b>5.6</b>	<b>4.5</b>
<b>Digit 3</b>	<b>31.6</b>	<b>22.5</b>	<b>9.3</b>	<b>6.7</b>	<b>1.9</b>	<b>2.3</b>
<b>Digit 4</b>	<b>0.8</b>	<b>2.0</b>	<b>10.4</b>	<b>14.4</b>	<b>20.5</b>	<b>25.0</b>
<b>Digit 5</b>	<b>0.8</b>	<b>2.5</b>	<b>8.9</b>	<b>10.6</b>	<b>18.3</b>	<b>18.2</b>
<b>Digit 6</b>	<b>1.7</b>	<b>3.0</b>	<b>7.9</b>	<b>8.9</b>	<b>12.1</b>	<b>18.2</b>
<b>Digit 7</b>	<b>0.8</b>	<b>3.0</b>	<b>6.9</b>	<b>7.2</b>	<b>9.0</b>	<b>9.1</b>
<b>Digit 8</b>	<b>1.7</b>	<b>3.7</b>	<b>5.5</b>	<b>6.7</b>	<b>7.2</b>	<b>6.8</b>
<b>Digit 9</b>	<b>2.5</b>	<b>3.0</b>	<b>5.4</b>	<b>4.9</b>	<b>5.4</b>	<b>4.5</b>
	-----	-----	-----	-----	-----	-----
<b>Data points:</b>	<b>237</b>	<b>4958</b>	<b>9090</b>	<b>4336</b>	<b>836</b>	<b>44</b>
<b>% Overall Data</b>	<b>1.2%</b>	<b>25.4%</b>	<b>46.6%</b>	<b>22.2%</b>	<b>4.3%</b>	<b>0.2%</b>

**Figure 4.75** Misguided Partition Distorts Development Pattern of U.S. Population Data

This empirical fact that other non-IPOT lenses lack vision, forensically confirms the theoretical reasoning given above about the uniqueness of IPOT partitioning.

The table in Fig. 4.75 is but one example of the severe distortion in the observed digital developmental picture under a misguided partition for the U.S. Census data on population centers, using the values 4, 40, 400, and so forth, as border points. No clear pattern whatsoever emerges here, and there is even a slight if confused trend of lower skewness as focus shifts to the right. Other non-IPOT partitions of the U.S. population data (not shown here) gave even worse results. Experimentations with other real-life logarithmic data sets as well as theoretical logarithmic distributions such as the Lognormal strongly confirmed all the results and conclusions drawn in this chapter.

Fortunately for the forensic data analyst, there are almost always three, four, or even more such IPOT sub-intervals. It is extremely rare for typical everyday data, including financial and accounting data, to be confined entirely within a single interval standing between two adjacent integral powers of ten (i.e. unity OOM). This fact makes this forensic development test quite feasible in almost all situations.

It is noted that insufficient as well as uneven data typically falls in the two extreme sub-intervals on the leftmost and on the rightmost sides of the entire IPOT partition. Assuming for example that the entire partition encompasses (1, 10000), then the leftmost sub-interval (1, 10) having the lowest values in the entire data set typically contains values mostly on its right part such as {7, 9, 4, 5, 9, 8, 9}, which constitutes an exact opposite of the Benford condition. The

rightmost sub-interval (1000, 10000), on the other hand, typically contains values mostly on its left part such as {2434, 1187, 3083}, namely severe digital inequality far more extreme than that of the Benford condition. This tendency is extremely typical in almost all data sets, since data points on the margins are highly sparse. All data sets tend to gravitate towards the center. This deficiency and unevenness of actual data within those two sub-intervals on the edges implies that leading digits there cannot fully express their true configuration even though these sub-intervals are properly standing between two adjacent IPOT points. To recap, typically most of the data in the leftmost sub-interval is within the right side of it where high digits lead, and most of the data in the rightmost sub-interval is within the left side of it where low digits lead, all because data on the margin always gravitates towards the overall center of the entire data set. In reality, this perceived bias is actually in complete harmony with the entire phenomenon of digital development, as it reinforces the overall trend of relatively stronger presence of high digits on the left, and relatively stronger presence of low digits on the right. It follows then that this extra push toward differentiation in digital configurations for the two extreme sub-intervals, enhances and reinforces that overall universal pattern of digital development in all random data types.

It might be argued that in order to detect digital development pattern in its purest form, these two extreme sub-intervals should be eliminated from the forensic analysis altogether prior to any calculations by default, even if more than 0.1% portion of data falls within any of them. Such an approach would utilize much lower threshold values for the four tests discussed in Chapter 43 — “Methods in Digital Development Pattern Detection”. This is so because empirical evidence from numerous real-life data sets shows that these two extreme intervals typically are responsible for a good portion of these four quantitative measures of development.

## DEVELOPMENT PATTERN MORE PREVALENT THAN BENFORD'S LAW ITSELF

---

---

Digital development or evolution from the left to the right for the random case is actually much more prevalent than even the Benford condition itself. In other words, this tendency shows itself not only in Benford data but also in almost all random and statistical data that are not logarithmic at all to begin with, such as corporate or governmental payroll accounting data, or U.S. Census County Area data. What accounts for this phenomenon? The answer is that the nature of random and statistical data (Benford or otherwise) in extreme generality is characterized by log-gradualism; that related log rarely starts nor ends abruptly around an initial or final value; and that it shows a marked curvature. Such log behavior in turn implies digital development at least to some extent. For non-Benford data, deviation from the logarithmic distribution is due typically to the fact that the range is not wide enough on the log-axis (low OOM), yet that overall log curvature is nearly universal, hence so is digital development. To recap, this assertion claims that even for data that is not logarithmic at all, a digital-dichotomy between two, three, or even four basic regions is almost always approximately being observed.

For a concrete example of this development pattern even for non-Benford data sets, census data on areas of all 3143 counties in the USA seen earlier in Chapter 45 is examined. First digits are {16.2, 10.0, 10.7, 15.8, 15.2, 10.4, 8.6, 7.1, 5.9} and the data is clearly not Benford at all. Nonetheless, a clear digital developmental pattern can be seen here. The table in Fig. 4.76 shows digital development of the county area data for the correct partition along adjacent integral powers of 10 points. A clear and almost consistent pattern of increasing favoritism towards low digits as focus shifts to the right is seen for this non-logarithmic data set.

The chart in Fig. 4.77 depicts related log histogram of U.S. county area. There is some vague resemblance to the Normal curve and the histogram starts and ends gradually on the log-axis itself without any abrupt spikes. It clearly shows two distinct regions: the one on the left rising, and one on the right falling. The absence of logarithmic behavior here can be clearly attributed to the fact that the

<b>Left Border Point</b>	<b>1</b>	<b>10</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>	<b>100,000</b>
<b>Right Border Point</b>	<b>10</b>	<b>100</b>	<b>1,000</b>	<b>10,000</b>	<b>100,000</b>	<b>1,000,000</b>
	=====	=====	=====	=====	=====	=====
<b>Digit 1</b>	<b>0.0</b>	<b>29.3</b>	<b>3.8</b>	<b>59.8</b>	<b>59.1</b>	<b>100.0</b>
<b>Digit 2</b>	<b>18.8</b>	<b>9.8</b>	<b>7.8</b>	<b>17.6</b>	<b>22.7</b>	<b>0.0</b>
<b>Digit 3</b>	<b>0.0</b>	<b>4.9</b>	<b>11.5</b>	<b>8.2</b>	<b>9.1</b>	<b>0.0</b>
<b>Digit 4</b>	<b>0.0</b>	<b>17.1</b>	<b>18.6</b>	<b>6.3</b>	<b>4.5</b>	<b>0.0</b>
<b>Digit 5</b>	<b>6.3</b>	<b>12.2</b>	<b>18.9</b>	<b>2.7</b>	<b>0.0</b>	<b>0.0</b>
<b>Digit 6</b>	<b>25.0</b>	<b>7.3</b>	<b>12.6</b>	<b>2.4</b>	<b>0.0</b>	<b>0.0</b>
<b>Digit 7</b>	<b>12.5</b>	<b>2.4</b>	<b>10.7</b>	<b>1.2</b>	<b>0.0</b>	<b>0.0</b>
<b>Digit 8</b>	<b>12.5</b>	<b>14.6</b>	<b>8.7</b>	<b>0.8</b>	<b>4.5</b>	<b>0.0</b>
<b>Digit 9</b>	<b>25.0</b>	<b>2.4</b>	<b>7.3</b>	<b>0.9</b>	<b>0.0</b>	<b>0.0</b>
	-----	-----	-----	-----	-----	-----
<b>Data points:</b>	<b>16</b>	<b>41</b>	<b>2408</b>	<b>655</b>	<b>22</b>	<b>1</b>
<b>% Overall Data</b>	<b>0.5%</b>	<b>1.3%</b>	<b>76.6%</b>	<b>20.8%</b>	<b>0.7%</b>	<b>0.03%</b>

Figure 4.76 Digital Development Pattern in Non-Logarithmic U.S. County Area Data

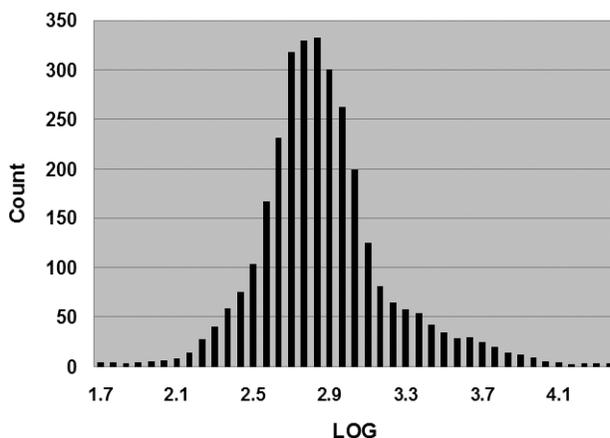


Figure 4.77 Related Log Histogram of Non-Logarithmic U.S. County Area Data

(bulk of the) spread on the log-axis is not nearly wide enough, approximately spanning only 1.4 log-axis units (since these two long whiskers of sparse data on the margins should be ignored in the determination of log spread and logarithmic behavior). Leading-digits inflection point here is around the log value of 2.8, which corresponds to the value of 630 for the data itself (calculated as  $10^{2.8}$ ). This value of 630 is approximately consistent with the digital development pattern seen in Fig. 4.76; as mini digit distributions are less skewed and milder than the

Benford condition to the left of this inflection point (approximated as 1000), and more extremely skewed to the right of it. The overall shape of related log curve seen in Fig. 4.77 is quite typical for almost all non-logarithmic random data; it is not very different from the typical shape of related log of logarithmic data; both are Normal-like, semi-circular-like, or triangular-like, and so forth; yet range on the log-axis is not wide enough for non-logarithmic data and thus the logarithmic cannot be realized.

## SUM-INVARIANT CHARACTERIZATION OF THE LAW (SUMMATION TEST)

---

---

Pieter Allaart's rigorous mathematical proof of equality of sums along digital lines is based on the premise that data or distribution falls exactly between two adjacent IPOT points, and hence data is of the **deterministic** flavor by default. Summation tests empirically performed on numerous real-life **random** data sets strongly refute the erroneous belief in the liberal extrapolation of such supposed equality to all real-life data. Proposition VI together with Allaart's assertion imply that sums along digital lines should be nearly equal in exponential growth series, so long as the series is long enough and closely stands between two IPOT points, adjacent or non-adjacent. Summation tests on abstract simulated exponential growth series representing deterministic data should then be contrasted with another abstract simulation representing the random process, namely numerical realizations via simulations of the Lognormal distributions having shape larger than, say, 1.0 (to insure logarithmic behavior). One such simulation of 35,000 realizations from the Lognormal with location parameter 7 and shape parameter 1.33 were conducted for the purpose of checking compliance with summation test. Resultant data set came out close to being perfectly logarithmic in the first-order sense, as well as in higher orders, FTD, and LTD senses as expected. The chart shown in Fig. 4.78 depicts sums declining along first digits, where sums for digits 1 and 9 differ by a factor of 7.2, strongly refuting that misguided belief in sum-equality for this generic random distribution. Sums along first-two digits (FTD) were also examined in this simulation, and results are shown in Fig. 4.79, where a downward trend is clearly visible, with sums differing roughly by a factor of 10 between those around 10 and those around 99.

We now turn our attention to summation tests for exponential growth series, representing the deterministic flavor. Figure 4.80 shows the first-order result for one such computer simulation for an exponential growth series starting with \$10 as the base, having 0.35% annual growth, and considered for exactly 3955 periods (years). Avoidance of extremely high values after so many growth periods

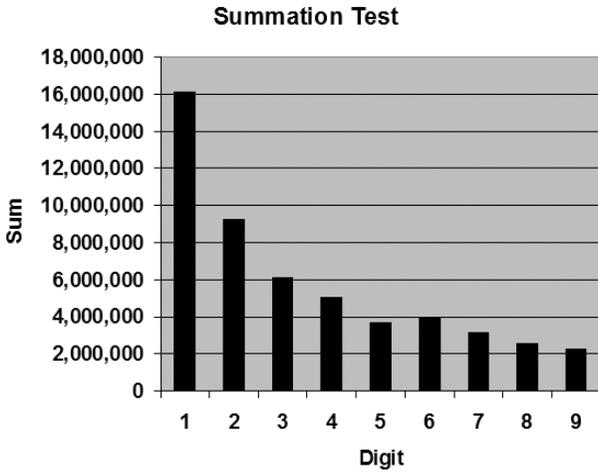


Figure 4.78 Sums Along First Digits, Lognormal — Location 7 and Shape 1.33

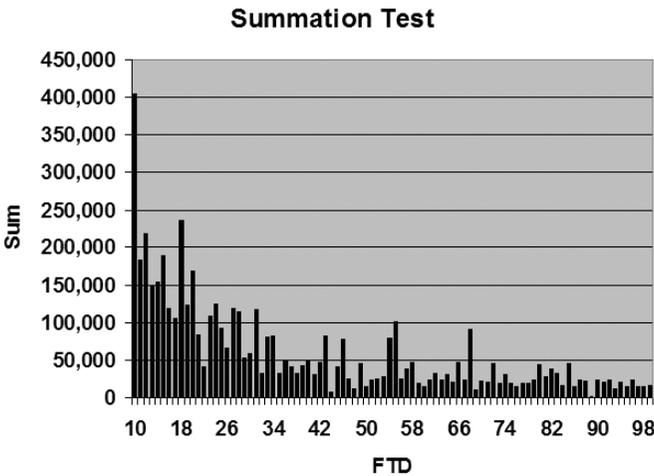


Figure 4.79 Sums Along First-Two Digits, Lognormal — Location 7 and Shape 1.33

motivated the choice of that very low growth rate of around a third of one percent. The choice of 3955 periods was motivated by the need to calibrate the series so as to make it start and end near IPOT points. The series starts at 10 and ends at 9993277.8, which is very near the IPOT  $10^7$ . As expected, the resultant series came out close to being perfectly logarithmic. Sums are quite equal here and summation test seems to be perfectly applicable in the deterministic case, strongly

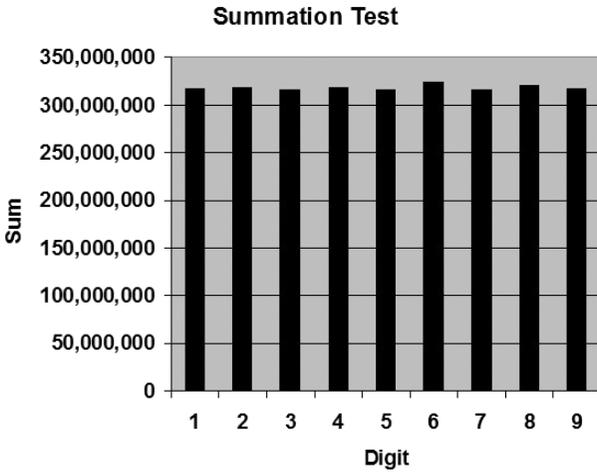


Figure 4.80 Sums along Digits, Exponential 0.35% Growth, Base \$10, 3955 Periods

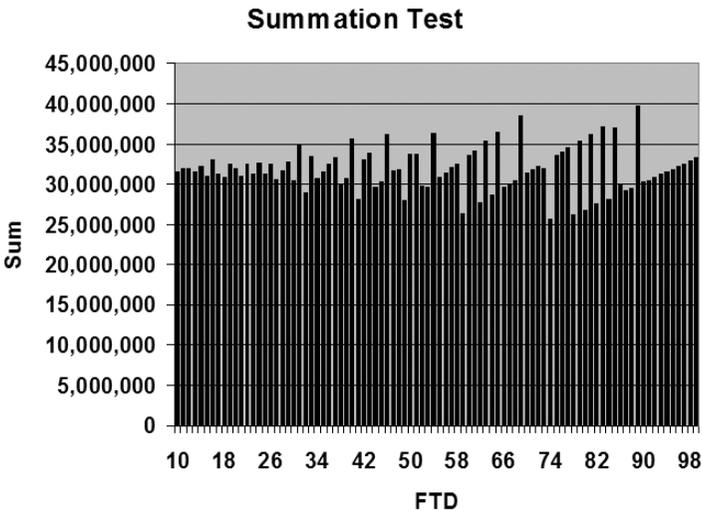


Figure 4.81 Sums along First-Two Digits, 0.35% Growth, Base \$10, 3955 Periods

confirming Allaart's assertion. A more refined examination at sums along first-two-digits line shows roughly steady and equal sums as seen in Fig. 4.81. Surely, computer simulation of the continuous  $k/x$  distribution defined over a range falling between IPOT values should also show equality of sums along digital lines, and more perfectly so. To recap, equality or inequality of sums along digital lines is

directly linked to the absence or presence, respectively, of any digital development pattern within the data. For the random flavor, the proportion of local sums for lower digits as compared with higher digits on IPOT sub-intervals increases as focus moves to the right, since the development is in the direction of increasing skewness towards the lower digits.

Let us formally prove Allaart's sum-invariant characterization in its first-order sense as well as in the general sense for higher orders for the particular deterministic case where  $k/x$  on  $(a, b)$  serves as the density, having a particular range such that  $a$  and  $b$  are two adjacent IPOT points of the particular  $(1, 10)$  case first, to be enlarged later to any other two adjacent IPOT values, such as  $(100, 1000)$ , as well as to non-adjacent IPOT values such as  $(1, 10000)$ . This result can be easily extrapolated for the discrete exponential growth series case where  $k/x$  serves as the approximating density. This alternative proof to the rigorous one given by Allaart derives the result via Proposition III and calculations of infinitesimal sums under  $k/x$  curve.

When a continuous random probability density distribution is considered, it is not possible to sum 'amounts' in the usual manner of adding discrete values on the histogram; there are none here. To better illustrate the point, one has to trace back the path we always take going from a histogram of discrete values to an abstract continuous density probability curve by way of dividing the height of each rectangle of unity width by the total number of data points in the entire data set. Hence we define a fixed imaginary integer  $N$  representing the number of all the values within some imaginary data set perfectly fitting the density curve in question, namely fitting  $k/x$  defined over  $(1, 10)$ .

The product of an infinitesimal tiny rectangular area under the curve times  $N$  should then represent the number of 'discrete values' falling within that infinitesimal sub-interval. If that product is then further multiplied by the value on the  $x$ -axis below the rectangular, it yields the 'sum' of all amounts within that rectangular area. In symbols:

Sum within mini sub-interval =  $x * N * [\text{mini area}]$ . Since it is necessary to sum those mini areas separately for each digit on  $(1, 2)$ , and on  $(2, 3)$ , ... , and on  $(9, 10)$ , the overall sum of all the mini areas' sums within any given digital-sub-interval is given by:

$$\text{Sum for digit } d = \sum x * N * [\text{mini area}] [\text{rectangles from } (d) \text{ to } (d+1)].$$

Turning it into a definite integral, we get:

$$\text{Sum for digit } d = \int x * N * f(x) * dx [\text{from } (d) \text{ to } (d+1)], \text{ or}$$

Sum for digit  $d = \int x^N k/x dx$  [from  $(d)$  to  $(d+1)$ ].

The  $x$  term cancels out, and we are left with:

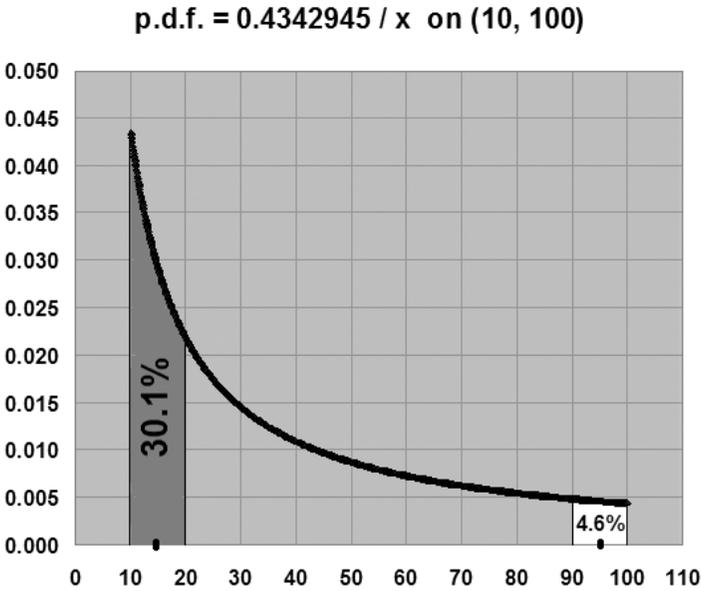
Sum for digit  $d = \int N k dx$  [from  $(d)$  to  $(d+1)$ ], or

Sum for digit  $d = N k [(d+1) - (d)]$ , or

Sum for digit  $d = N k$

Namely the same constant value  $N k$  for each of the nine digits, hence summation equality between the digits in this particular case is proven. Pure mathematicians may question the basis of this proof and call for further scrutiny, but for open-minded statisticians the setup of this proof surely would resonate as something quite familiar. In the discrete case the idea of summation is always employed in the definition of the average, as in  $[Avg] = [Sum] / [Number\ of\ Values]$ . Here, too, the expression  $\int x^N f(x) dx$  corresponds to the generic definition of the average  $\int x f(x) dx$  [over entire range] in mathematical statistics except that extra  $N$  term. But since  $N$  does indeed represent [Number of Values], the expression  $\int x^N f(x) dx$  directly signifies sum!

Two essential features provide for this equality. First it's the constant integrand after the cancellation of the  $x$  term, which is due to that very unique  $f(x)$  form of  $k/x$ ; no other functional form would do! **This fact may be viewed as the source of the uniqueness of  $k/x$  in the context of Benford's Law, and which is what is behind the temptation to characterize Benford's Law in terms of sum invariance. Yet this is true only for the deterministic flavor and it ignores the random flavor altogether and the consequences of Proposition VII.** The second essential feature is the equality of length on the  $x$ -axis in the limits of integration for the various digits [since  $(d+1)$  minus  $(d)$  is 1, thus it is a constant for all the digits]. Since these two features are present just the same for all other  $k/x$  cases with adjacent IPOT points such as  $(100, 1000)$  and so forth, the proof can then be enlarged to include them as well. For example, in  $k/x$  on  $(100, 1000)$  the limits of integration for each leading digit  $d$  are from  $100*(d)$  to  $100*(d+1)$ , therefore the length on the  $x$ -axis is the same for all digits, while the value of the integrand is also constant here just the same. Even in the case where  $k/x$  is defined over non-adjacent IPOT values, such as  $(1, 100)$ , sum equality still holds as more than one set of limits of integration are considered and the terms involving  $(d)$  and  $(d+1)$  all yield constant differences. For the  $(1, 100)$  case, the  $d$ -terms cancellations are within  $(d+1) - (d)$  [belonging to the sub-interval  $(1, 10)$ ] as well as within  $10*(d+1) - 10*(d)$  [belonging to the sub-interval  $(10, 100)$ ], and

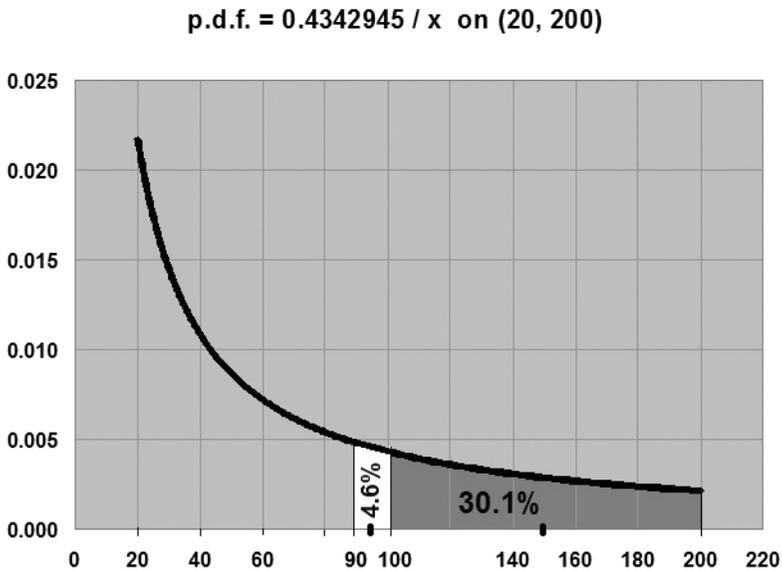


**Figure 4.82** Equality of Sums along Digits in the Case  $k/x$  Defined Over (10, 100)

together the two differences are equal to the constant  $11 \cdot N \cdot k$ , the same for all the digits, and thus equality holds here just as well.

To illustrate, let us calculate sums for the case of  $0.43429/x$  over (10, 100) as depicted in Fig. 4.82. Sum for digit 1 is calculated here as  $N \int x(k/x)dx$  [from 10 to 20], or  $N \cdot k \cdot 10$ , which is  $N \cdot 4.3429$ . For digit 9 sum is similarly calculated as  $N \int x(k/x)dx$  [from 90 to 100], or  $N \cdot k \cdot 10$ , which is also  $N \cdot 4.3429$ . Sums are equal here for digits 1 and 9, as well as for all digits. Digit 9 stands around  $x = 95$  territory, while digit 1 stands around  $x = 15$  territory, therefore on this account digit 9 comes with **6.33-fold territorial advantage** over digit 1 (calculated as  $95/15$ ). But digit 1 comes with **6.58-fold Benfordian advantage** over digit 9, calculated from the logarithmic proportions of  $(30.1)/(4.6)$ , hence it all cancels out exactly. [Taking the midpoint of each territory is only an approximate way to express territorial location where curve is sloping, hence the slight difference in factors].

The above proof is not valid, though, for cases where  $k/x$  is defined over an integral exponent difference having non-IPOT edges, such as (20, 20000), nor even for cases where  $k/x$  is defined over a unity exponent difference having non-IPOT edges, such as (20, 200), since limits of integration are not of equal length



**Figure 4.83** Inequality of Sums along Digits in the Case  $k/x$  Defined Over  $(20, 200)$

for all digits, and also because classic relative positions of digits are reversed here. For example, in the perfectly logarithmic case of  $k/x$  over  $(20, 200)$ , limits of integration for digit 1 span the length of 100 (from 100 to 200), while for digits 2 to 9 they span only the length of 10, hence sums vary considerably. Moreover, there is a profound shift here in the order of the outlay of the digits on the x-axis. Figure 4.83 clearly illustrates this glaring and obvious counter-example to the mistaken ‘extrapolation’ of equality of sums to all the wrong places, the myth of it being a universal property of the Benford condition, an error that has been frequently committed in the literature. Density  $k/x$  is defined over  $(20, 200)$  and it is perfectly logarithmic. Yet, the ‘natural’ order of digits 2 to 9 is inverted as they all fall back behind digit 1, standing on a much less valued territory on the left where amounts are lower. Digit 1 is standing proudly far ahead of the pack in front of everybody else to the right, and thus earns by itself more sum than the combined sums of all the other eight digits! This is so not only by virtue of possessing territory of the best quality in the whole range, but also due to its quite large portion of overall data as per the law (30.1%).

Let us actually calculate sums in the case of  $0.43429/x$  over  $(20, 200)$ . Sum for digit 1 is calculated here as  $N \int x(k/x)dx$  [from 100 to 200], or  $N*k*100$ , which

is  $N \cdot 43.429$ . For digit 9, sum is calculated as  $N \int x(k/x) dx$  [from 90 to 100], or  $N \cdot k \cdot 10$ , which is  $N \cdot 4.3429$ . That is a whopping tenfold sum advantage for digit 1 as compared to digit 9! Nothing is equal here! Why such an uneven result? The straightforward answer is that digit 9 stands roughly around  $x = 95$  territory, while digit 1 stands around  $x = 150$  territory, therefore on this account alone digit 1 comes with **1.58-fold territorial advantage** over digit 9 (150/95). In addition, digit 1 comes with **6.58-fold Benfordian advantage** over digit 9, calculated from the logarithmic proportions of  $(30.1)/(4.6)$ . Taking stock now, we multiply these two factors to arrive at the overall digit 1 advantage factor of  $[1.58] \cdot [6.58] = 10.39$ , which is comfortably close to the actual (exact) tenfold factor.

The above proof can be easily extended to higher orders; hence all results here are true in the general sense of the law. This can be clearly demonstrated by pointing out to the much shorter cycles of higher orders which are all contained in any range between IPOT points. For example, (10, 20) is a full cycle for the second order and (37, 38) is a full cycle for the third order. Therefore, the outline of the above proof easily carries over to the same summation equality along first-two-digits (FTD) line as well, where instead of integrating on the intervals  $(d, d+1)$ , we integrate on much smaller sub-intervals to obtain the same sum equality. For example, for  $k/x$  defined over (1, 10), sums along FTD lines are calculated on (1.0, 1.1), (1.1, 1.2), (1.2, 1.3), ... , (9.7, 9.8), (9.8, 9.9), (9.9, 10), and with all 90 sub-intervals spanning the same length of limits of integration, resulting in the same uniformity of sums seen earlier.

In spite of the widespread misguided and erroneous application of Allaart's sum-invariant characterization of BL in accounting and finance, and in spite of its very restricted condition requiring data to stand exactly between two IPOT points and to fall off exactly as in the  $k/x$  distribution, its theoretical consequences lead to a more profound understanding of the meaning of 'logarithmic-ness' as shall be discussed in Section 7. Sum-invariance of data on (1, 10), for example, requires data to continuously fall as it is not compatible with a flat or rising histogram. Moreover, it requires the data to fall at a certain very particular rate of sharpness while being already properly aligned and coordinated along the  $x$ -axis between IPOT values as per its defined range.

Finally, digital development was shown to exist universally for all random data, Benford or otherwise. Therefore sums for low digits by far outweigh sums for high digits even for non-logarithmic data, and this inequality therefore represents a universal property for all random data types. As one particular real-life

manifestation of this **sum-variant universal property of random data**, U.S. County Area data seen earlier shall be considered. For this decisively non-logarithmic random data set, sum along digit 1 is about fourfold the sum along digit 9. The vector of nine sums along first digits is: {870662, 441400, 356916, 425720, 349850, 300924, 253092, 310737, 222605}, hence, sum along digit 1 (i.e. 870662) is about four times the sum along digit 9 (i.e. 222605).

**This page intentionally left blank**

## **Section 5**

# **BENFORD'S LAW IN THE PHYSICAL SCIENCES**

**This page intentionally left blank**

## MOTHER NATURE BUILDS AND DESTROYS WITH DIGITS IN MIND

---

Mother Nature is always quite busy, simultaneously forming planets, stars, galaxies, rivers, cities, and towns. Yet, in spite of her hurried and difficult schedule she is very picky when it comes to quantitative style and so she takes her time, deliberately creating them her way. Even in her rare moments of anger, when she rattles planets with earthquakes and erupts volcanoes, even when she rages and spectacularly lightens the sky with supernova explosions, smashing large stars to smithereens, she still exercises some self-control and pauses a moment to calculate resultant relative quantities carefully, so as to create or destroy in her own particular way. She greatly favors and sympathizes with the small and the weak and therefore she creates many of them while disliking and suspecting the big and the powerful, producing very few of them.

She has only very recently learnt of digits and numbers, logarithms and mantissa, histograms and densities — invented by those troublesome humans — and she adamantly claims to avoid utilizing all of these strange abstractions in her daily work and constructions. Yet, curiously her creation always exhibits a great deal of digital pattern and order, in spite of her constant denials. So much so that now she finally admits to it, but still insists that when it comes to digits she does it subconsciously, and that she doesn't really know why.

This section is devoted to analyzing the way Mother Nature indirectly forms numbers and digits, and the enormous prevalence of Benford's Law in physics, chemistry, astronomy, geology, biology, economics, and other physical and social sciences. One of the key reasons that the logarithmic is frequently found in such disciplines is that many such measures and variables are Lognormally distributed (and hopefully in our context with high value of the shape parameter in order to

obtain logarithmic behavior). Quoting from [Wikipedia](#): The Lognormal distribution is important in the description of natural phenomena, such as:

**In biology:** Measures of size of living tissue (length, skin area, weight); The length of inert appendages (hair, claws, nails, teeth) of biological specimens; certain physiological measurements, such as blood pressure of adult humans.

**In hydrology:** The Lognormal distribution is used to analyze extreme values of such variables as monthly and annual maximum values of daily rainfall and river discharge volumes.

**In social sciences and demographics:** Income of 97% – 99% of the population is distributed Lognormally.

**In technology:** The Lognormal distribution is often used in reliability analysis to model times to repair a maintainable system. In wireless communication, the local mean power expressed in logarithmic values, such as dB or neper, has a Normal distribution.

Yet, curiosity prompts one to ask why so many physical phenomena are Lognormally distributed in the first place. The following chapters will explore these issues further, and attempt to give a reasonable answer to this question in some particular cases.

## QUANTUM MECHANICS, THERMODYNAMICS, AND BENFORD'S LAW

---

Physics has given rise to three essential statistical distributions that are widely used in quantum mechanics and thermodynamics, having either the same form of the exponential distribution or approximately so. The probability density function of the exponential is  $\text{PDF}(X) = \lambda e^{-\lambda X}$ ,  $X > 0$  and  $\lambda > 0$ . Overall leading-digits distribution of the exponential distribution is nearly but not exactly logarithmic, and this is true regardless of  $\lambda$  parametrical value. Individual first digits oscillate tightly and closely above and below their  $\text{LOG}(1+1/d)$  centers in unsynchronized cycles depending on exact parametrical  $\lambda$  value. In other words, digits deviate slightly from the logarithmic, as a function of parameter  $\lambda$ . To the extent that the exponential is close enough to the logarithmic, all this implies that Benford's Law is highly relevant and prevalent in those fields of physics.

The three relevant distributions are: Boltzmann-Gibbs, Fermi-Dirac, and Bose-Einstein distributions. Lijing Shao and Bo-Qiang Ma (2010) at Peking University examined the relationship between these distributions and Benford's Law and concluded that the logarithmic represents a general pattern in statistical physics. Moreover, they conjectured that Benford's Law itself might be (or might hint at) a truly profound and fundamental law in nature in general, and not only in statistical physics.

The Boltzmann-Gibbs distribution in thermodynamics relates to the behavior of particles according to classical physics where quantum effects can be ignored. It explains macro properties of matter such as temperature, pressure and volume, by way of analyzing their micro (atomic/molecular) chance behavior. It plays a fundamental role throughout statistical physics. Ludwig Boltzmann is credited for having begun this branch of statistical mechanics with a basic paper written in 1884. His fundamental work was taken up again, re-written, and expanded in a classical treatise by Josiah Willard Gibbs.

The expression of Boltzmann-Gibbs distribution is  $F_{\text{BG}}(E) = \beta e^{-\beta E}$ , where  $\beta = 1/kT$ ,  $T$  is the temperature,  $k$  is the Boltzmann constant, and independent

variable  $E$  represents the system's energy. Being that the distribution itself is exactly exponential in form, with digits oscillating closely around their average values of the Benford proportions, the connection to Benford's Law is immediate. Moreover, if multiple values of temperature  $T$  are considered, then that single density representing the aggregate of all the individual densities with varying  $T$  values gets very close to the logarithmic (due to the averaging and canceling effects on these cycles), by far closer than any single distribution with a singular  $T$  parameter could aspire to be. In fact if  $T$  can be considered as a variable (i.e. distribution) then this process can be thought of as a two-sequence chain. If in turn  $T$  can be considered a logarithmic distribution or close to it, then by the second chain conjecture  $F_{BG}(E)$  is logarithmic or extremely close to it. A more detailed account will be given in Chapters 102, 103, and 107 about the exponential distribution, such density aggregates as chains, and their near-logarithmic behavior.

The Fermi-Dirac distribution was discovered independently by both Enrico Fermi in Italy and Paul Dirac in England at around the same time in 1926. It relates to electrons' behavior and spans the fields of quantum mechanics and thermodynamics. The expression of Fermi-Dirac distribution is  $F_{FD}(E) = [\beta/\ln(2)] * [1/(e^{\beta E} + 1)]$ . This is not exactly an exponential distribution due to the extra  $(+1)$  term, but it is very nearly so since the term is marginal in the grand scheme of things. Empirical examinations of the distribution show that digits oscillate nicely and closely around their logarithmic centers, almost exactly as in the case of the exponential distribution.

The Bose-Einstein distribution was first suggested by Bose in 1924. Bose, an obscure physicist from India as far from mainstream physics at the time as one could imagine, sent his article to Einstein after a rejection. Einstein immediately recognized its significance and extended it to atoms. Statistical physics was certainly Einstein's early love. He had written papers on the statistical basis of thermodynamics even before his 1905 paper on relativity and has contributed significantly to the field. But Einstein the statistical physicist was ultimately eclipsed by Einstein the relativist. Einstein was highly enthusiastic about applying statistical theory to physical systems as consequences of scientific laws, but objected and detested utilizing statistics in the very formulations of these fundamental physical laws, coining the motto "God does not play dice."

Mark Haw (2005) of the University of Edinburgh writes: "Einstein's role in demystifying Brownian motion [and in demonstrating that atoms are real, not merely abstract entities serving as an analogy] was pivotal in this ongoing

revolution. In developing the first testable theory that linked statistical mechanics — with its invisible “atoms” and mechanical analogies — to observable reality, Einstein acted as a gateway. Through this gateway, years of confused observations could be turned into the solid results of Perrin, and from these could grow a new, proven world view with statistics at its heart. From our more distant perspective, it is clear that the [Einstein’s] Brownian-motion papers of 1905 had just as much influence on science as did relativity or light quanta. Brownian-motion was just a slower, subtler revolution, not a headlong charge.”

The expression of Bose-Einstein distribution is  $f_{BE}(E) = 1/(e^{\beta E} - 1)$  where the equality sign should be replaced by a proportional sign. This is not exactly an exponential distribution due to the extra  $(-1)$  term, but it is very nearly so. Empirical examinations show that digits oscillate nicely and closely around their logarithmic centers, almost exactly as in the case of an exponential distribution.

## CHEMISTRY, RANDOM LINEAR COMBINATIONS, AND BENFORD'S LAW

---

Buck (1992) investigated leading digits of alpha decay half-lives. Values of half-lives of radioactive materials have been accumulated throughout the 20th century, and vary over many orders of magnitude, making the set a prime candidate for logarithmic behavior, in spite of its small size. Buck found a very good agreement with the law for all available 477 measured values of alpha half-lives, with first digits as {29.6, 17.8, 11.7, 10.5, 9.9, 4.8, 5.3, 5.2, 5.2} and the implied very low SSD value of 9.7. Exceptionally large order of magnitude of variation here must have compensated for the scarcity of values, resulting in such strong conformity with the law. The shortest half-life of about 0.299 microseconds ( $0.299 \times 10^{-6}$  second) is that of an extremely unstable isotope of Polonium, 212 Po. The longest half-life of  $7 \times 10^{+15}$  years is that of a very stable isotope of Samarium, 148 Sm. Putting both periods on the same time scale footing (days in this example) yields an exceedingly wide range in terms of order of magnitude of ( $3.5 \times 10^{-12}$ ,  $2.6 \times 10^{+18}$ ) days! No other data set in this book comes even close to having such an unusually large value ( $\approx 30$ ) of order of magnitude! Although Buck's data focuses solely on half-lives of unhindered alpha decays, he predicts similar digital results for radioactive decays in general, regardless of the precise mode of disintegration. Buck also suggests that a closer look at the details of half-life probability distributions themselves might yield information about the physical processes involved in relation to digital behavior (perhaps in a similar manner seen earlier regarding thermodynamics, with either the exponential itself or other exponential-like distributions serving as the density). The processes of all such decays are the combined effects of the short-range nuclear attractive strong forces and the long-range repulsive electromagnetic Coulomb forces, all of which give rise to a potential well and barrier able to keep an alpha particle bound inside the nucleus for some (probabilistic) length of time.

The **atomic weight** values of the **elements** within the Periodic Table are not logarithmic, although digits 1 and 2 predominate, taking 73% of total leadership altogether. First-digits distribution is {35.9, 36.8, 4.3, 3.4, 5.1, 3.4, 3.4, 3.4, 4.3}

yielding the very large SSD value of 536.2. A closer look at the Periodic Table shows that from Hydrogen with a weight of 1.008 to Technetium with a weight of 98, all digits 1 to 9 get roughly fair chances to compete for leadership. Then, for the 36 elements from Ruthenium (101.07) to Gold (196.97) first digit is always 1. For the rest of the 39 elements, from Mercury (200.59) to the heaviest of them all, Ununoctium (294), first digit is always 2. In any case, the deviation of the Periodic Table from Benford is due not only to small data size, but more importantly to the small order of magnitude, calculated as  $\text{Log}(294) - \text{Log}(1)$ , namely 2.5, which is insufficient. Could Benford's Law then be observed perhaps for the **molecular mass of compounds**?

A list of 2175 common chemical compounds in use worldwide is given in the website <http://www.convertunits.com/compounds/Z/>. The selection of compounds was not made following any strict criteria in some systematic manner, but rather simply by gathering information from many different relevant sources, and following the suggestions of users and chemists. The impressive variety in the list makes it appear to be a good and fair representative of any proper collection of chemical compounds in use for the purpose of testing for compliance with Benford's Law. It includes seemingly totally unrelated chemicals, such as those used in heavy industry, the pharmaceutical industry, the food industry, metallurgical plants, as well as many naturally occurring compounds along with synthetic ones. Certainly, no attention whatsoever was paid to molar (molecular) mass in the process of selecting and compiling this list. The molar mass is the combined/total mass of all the elements within the molecule. For example, the molar mass of the  $\text{H}_2\text{O}$  water molecule is 18.01528, having two hydrogen elements of 1.00794 weight each, and one oxygen element of 15.9994.

As an example, the following five compounds are part of the long list of 2175 molecules:

Silicon Iodide – $\text{SiI}_4$	molar mass = 535.70338 g/mol
Copper Nitrate – $\text{Cu}(\text{NO}_3)_2$	molar mass = 187.5558 g/mol
Glucose – $\text{C}_6\text{H}_{12}\text{O}_6$	molar mass = 180.15588 g/mol
Hydrogen Peroxide – $\text{H}_2\text{O}_2$	molar mass = 34.01468 g/mol
Iron(II) Bromide – $\text{FeBr}_2$	molar mass = 215.653 g/mol

The respective first- and second-digit distributions of the molar mass in this list are:

{31.9, 25.2, 16.1, 8.4, 5.7, 4.3, 2.9, 3.2, 2.3}  
 {11.2, 9.9, 11.1, 10.1, 10.3, 10.7, 9.2, 9.0, 9.8, 8.6}

This is quite close to the logarithmic and constitutes one quite remarkable result! SSD for the first digits is somewhat moderate at 102.9. Curiously, Mother Nature's way of playing that Lego-like game, combining elements to form compounds, somehow yielded the logarithmic distribution once again in the physical world. One would certainly suppose that Mother Nature herself paid no attention whatsoever to resultant molar mass in sticking those elements together — well, at least not directly — as she simply follows her own complicated quantum mechanics dictate regarding electronic shell configurations, offering elements and compounds that heartbreaking and morally dubious choice of either sharing or stealing electrons from their loyal neighbors. Surely, she hasn't paid even scant attention to digits, the invented fantasies of those creative but troublesome humans, nor has she read Benford's 1938 article purporting to observe some kind of exact proportions between those imaginary abstractions. She would be quite surprised to learn that she has done exactly that (albeit unintentionally), following and obeying what he had predicted! But why should molar mass be logarithmic?

Curiosity compels one to compare this remarkable result with a totally random selection, combination, and mixing from the Period Table. Would such blind and haphazard combination yield similar digital results? Monte Carlo computer simulations of just such a scheme were performed, with random selection from the first 35 elements in the Period Table, ranging from simple hydrogen all the way to the element Bromine with its mass of 79.904 as the heaviest one in this simulation scheme. Bromine was arbitrarily chosen so as to avoid the heaviest elements which are rarely in use in industry and human activities. The simulations aimed at randomly building a molecule out of two or three elements (determined by a flip of a coin). The number of atoms for each element chosen was randomly determined to be anywhere from a single atom up to five repetitions. Symbolically the simulation scheme for the molar weight of the invented molecule was as follows:

$$[\text{dice of } 5] * \text{element1} + [\text{dice of } 5] * \text{element2} + \text{COIN} * [\text{dice of } 5] * \text{element3}$$

The first simulated element within this factitious or virtual molecule is then chosen as in the equal discrete uniform distribution  $\{1 \text{ to } 35\}$ , and its frequency within the molecule is simulated separately from the discrete uniform  $\{1 \text{ to } 5\}$ . The second element is chosen in likewise manner, independently, and with possible duplication of elements and quantities. The third and final element is likewise simulated, but only if a coin is flipped showing a head [valued as 1]. Otherwise, when a tail is obtained [valued as 0], it is aborted and the molecule is

being built just from the previous two selections. Thus the highest number of total elements a molecule can hope to have in such simulations is  $3 \times 5$ , while the least fortunate compound would have just  $2 \times 1$  elements. The fact that most molecules in the list of real common compounds actually consist of two to three distinct elements, and that each element usually appears no more than five times within a given compound, has provided the motivation for the above parameters and structure in the computer simulations. The first- and second-digit distributions results of 10,000 such computer simulated molecules came out very close to those of the list of real chemical compounds above, and are respectively:

{31.4, 25.0, 16.2, 8.6, 5.4, 3.2, 3.1, 3.6, 3.5}

{11.7, 11.1, 11.2, 10.7, 10.2, 9.4, 9.9, 8.7, 8.6, 8.5}

Furthermore, the two adjusted histograms of the data sets themselves (real and simulated), of the type that puts both data sets on an equal footing by listing percent of overall total (2175 and 10,000) instead of counts, are remarkably similar, as shown in Fig. 5.1. This result strongly suggests that nature's way of gluing elements together into molecules is done in an almost totally random manner with regards to molar mass. Consequently, the process of the formation of molar mass

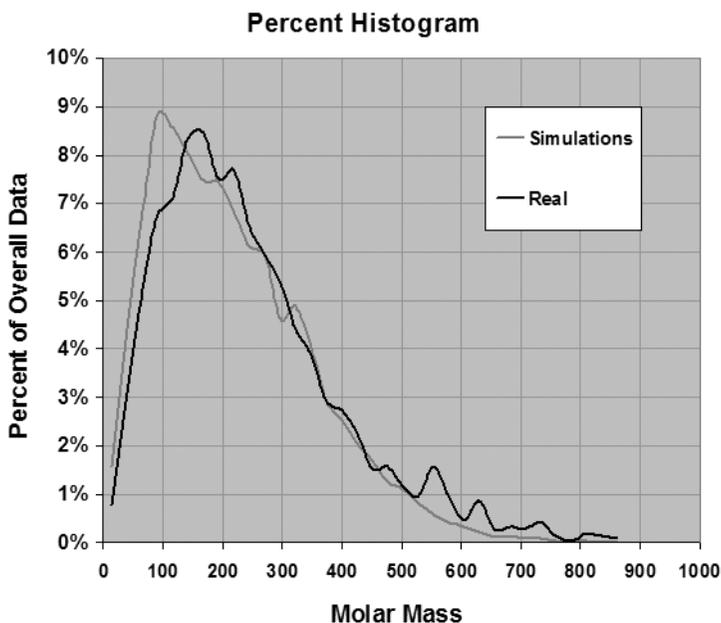


Figure 5.1 Histogram of Simulated and Real Data on Molar Mass of Compounds

in compounds can be viewed as statistical in nature and studied as such. Clearly there are chemical/physical limitations as to how elements can be combined; there are exact laws, regularities, and limits on what is allowed and what is not allowed, yet apparently atomic/molar mass does not play any direct role in how compounds are being formed.

An interesting observation here is made following a close scrutiny of the above digital results for the real and simulated molar mass, namely that while this element-gluing process comes with a unique digital signature in the first order sense, there exists no such authentic signature in the second order sense. Chapter 71 of 'The Near Indestructibility of Higher Order Distributions' almost certainly implies that variations in the second order here are random in nature, fluctuating around the theoretical unconditional proportions, while the (nearly) steady and particular digital signature of the first order is structural in nature, pertaining to this particular chemical process. If the list of invited honorary guests to the Bolshoi Theatre in Moscow contains {Smirnov, Ivanov, Sokolov, Popov, Kozlov, Novikov, Morozov, Solovyov}, then the designation part on the right of 'ov' [akin to the universal second order] does not truly identify the person, only the letters on the left do [akin to the particular first order].

It is important to note how the backdrop of the structure of the Periodic Table itself plays an essential role here and strongly affects real and simulated results. Dmitri Mendeleev's table contains a large variety of distinct weight values, all of which increase monotonically, thus guaranteeing a unique weight for each element. Had simulations or real-life chemical compounds been based on a hypothetical table of 111 elements having just three distinct weight values, with many elements having identical mass values, there would have been much less variability here and Benford's Law wouldn't have been observed. On the other hand, had simulations been conducted with a freer hand in choosing a much larger number of distinct elements, as well as allowing unrestricted repetitions within one single molecule, such as 34 Carbons, 124 Chlorines, and so forth, then variability would have been sufficient and Benford's Law could have been observed even with a table having very few distinct weight values. In such tug-of-war situations between a lack of variety within the table and the liberal manner allowed in constructing a molecule, Benford wins or loses depending on the relative pull each side exerts.

In the above simulations of chemical compounds, Bromine and its mass of 79.904 was arbitrarily chosen as the heaviest element, and hydrogen and its mass of 1.00794 as the lightest. The order of magnitude of the 'element list' (i.e. the reduced Periodic Table) is  $\text{LOG}(79.904/1.00794)$  or 1.9. Actual resultant data

should have a much higher value of OOM since simulations allow up to five repetitions of three elements. In actual simulations OOM came out approximately 2.6, although its full potential is slightly higher considering two extreme hypothetical molecular selections which did not actually occur in the simulations due to their rarity. The first is the quintuple selection of three very heavy elements, say Arsenic, Selenium, and Bromine, having atomic mass of 74.921, 78.960, 79.904 respectively, which potentially results in a molar mass of  $5 \times 74.921 + 5 \times 78.960 + 5 \times 79.904 = 1168.928$ . [And even heavier selection would be 15 Bromine atoms.] The other extreme is choosing just two light hydrogen atoms yielding a molar mass of  $2 \times 1.00794 = 2.01588$ . Calculating potential OOM gives  $\text{LOG}(1168.928 / 2.01588)$ , or 2.8, which is sufficiently large to yield digits distribution comfortably close to the logarithmic. Yet, according to the OOM qualification, the last calculation leading to the value 2.8 is quite misleading and incorrect since these rare molecular constructions are surely to be considered outliers. Furthermore, even the observed OOM value of 2.6 in resultant data should be reduced slightly due to the need to calculate variability solely on the bulk of the data omitting roughly 10% on either side. Indeed only OMV measure should be applied in the determination and the expectation of seeing logarithmic behavior in any dataset.

A remarkable correspondence and analogy to the molar mass of chemical compounds is found in accounting data relating to company revenues. A molar mass is the mass of the entire molecule made of multiple elements. A purchasing bill typically contains multiple prices of the various items bought. The mass of each element is distinct and is found in the Periodic Table. The company's price list of its many products on sale similarly comes with distinct prices, or nearly so. A molecule is typically made of two to three elements, and at times even four or five. Similarly, the bill of a large purchase typically consists of two, three, or more distinct products. Repeating elements are typical, as in the water molecule  $\text{H}_2\text{O}$  where hydrogen repeats but oxygen appears only once. Similarly, in a typical purchase, two, three, or more quantities of one popular product are bought in the same order, while only one quantity perhaps of another less popular product is bought and included in the bill. In this context, the individual weight of an element in the Periodic Table is analogous to the price of an item on sale in the company's list of prices; and the molar mass is analogous to the total bill paid by the shopper. Accounting revenue data are (empirically) known to obey Benford's Law fairly well. Both revenue data and chemical molar mass data are simply two particular manifestations of the much wider and general principle in Benford's Law regarding

**Random Linear Combinations (RLC).** Surely there are numerous other real-life data sets and processes having the same underlying mathematical structure that comes under the protective umbrella of Random Linear Combinations and are therefore logarithmic as well. The term 'Random Linear Combination' as defined and referred to in this book pertains to a random process obtained by repeated additions of any multiples of values in a given data set (a very **long list** of numbers essentially), including a single multiple of it without any addition terms. In symbols:  $RLC = N_1 * LIST + N_2 * LIST + N_3 * LIST + \dots + N_L * LIST$ , with the possibility that each  $N_i$  is also some random number such as a dice or a numbered coin, and typically thought of as a very **short list** of integral numbers distributed as in the discrete uniform (or as in other discrete distributions). Even the value  $L$  — the numbers of terms to be added — may be random. This definition excludes the possibilities of power terms such as  $LIST * LIST$ ,  $2^{LIST}$ ,  $LIST^{13}$ , division as in  $1 / LIST$ , as well as all other non-linear combinations, all of which do not correspond to supermarket's bills or actual weights of molecules.

## BENFORD'S LAW AND THE SET OF ALL PHYSICAL CONSTANTS

---

---

The set of some 337 physical constants assuming the metric system and other scales can be found at <http://physics.nist.gov/cuu/Constants/Table/allascii.txt>. The list includes items such as Planck and Boltzmann constants, Bohr radius, joule-atomic mass unit relationship constant, constants relating to gravitational, electromagnetic, and quantum forces, as well as others in physics and chemistry. Remarkably, its first-digits distribution is {34.7, 19.0, 8.9, 8.3, 8.3, 7.1, 3.3, 5.1, 5.3}, which is quite close to the logarithmic in spite of the severe scarcity of values here. This interesting result is unique in the field of Benford's Law, standing out apart from many other digital results. It cannot be explained as being simply another manifestation of the logarithmic in the natural world, since the set is not at all about a single physical issue or phenomenon but rather about the format and regularities governing **all** known phenomena. Nor can it be explained in terms of other logarithmic models to be proposed in the next chapters regarding the physical manifestation of Benford's Law. Surely it cannot be modeled on some mysterious multiplication processes or as a distribution of many invisible and hidden distributions. It surely lends the law some more weight and mystique. It is also interesting to note that to this short list of numbers we owe our existence in some sense. The particular values within this data set, or rather the relative magnitudes between these values, gave rise to **the anthropic principle in astrophysics and cosmology**, namely the observation that the physical universe must be compatible with the conscious life that observes and records it, noting the remarkable fact that the universe's fundamental constants just happen to fall within that very narrow range of relative values which makes life possible. For example, had the gravitational constant been just a bit lower, stars would have insufficient pressure to overcome the barrier of the repulsive electromagnetic force needed to start thermonuclear fusion in pressing hydrogen atoms together into helium, and therefore wouldn't shine nor produce heavier elements like carbon, iron, and

oxygen. Had the gravitational constant been just a bit higher, stars would burn too fast, thus quickly depleting their fuel and rendering life impossible.

And what would leading digits of physical constants look like had we calculated them all assuming other scales such as feet, pounds, hours, and so forth? The answer is that for the most part scale changes would not affect leading-digits distribution by very much, although a slight change would indeed be noticeable. Had a change of scales been shown to produce some dramatic deviation from that near-logarithmic condition we currently have under our metric system, one might then conclude or speculate that the present set of units of measurements has some intimate connection with the real world and therefore unique! Yet, our scales are not of such special quality. The arbitrary choices of our scales and units in use came about for historical reasons, much of it occurring during and after the French Revolution with regards to the Meter and the Kilogram under the custody of the French Academy of Sciences, with the active support of Napoleon Bonaparte.

## MCLT AS AN EXPLANATION FOR SINGLE-ISSUE PHYSICAL PHENOMENON

---

---

The merging of two findings points to a plausible explanation for the logarithmic behavior in some cases of single-issue physical data sets. (I) The understanding that the generic shape of histograms of typical logarithmic data sets is skewed with a tail to the right and the implied skewness in relative quantities. (II) The observation that random multiplication processes result in the same relative quantity structure where the small outnumbers the big. The strong connection between these two findings shall now be demonstrated and then further utilized.

As mentioned in the chapter on Hill's super distribution, his model cannot be applied as an explanation for single-issue physical data. Instead, a plausible explanation could be proposed in the argument that such physical manifestations of the law are somehow the cumulative effects of numerous multiplicative random factors, which leads to the (logarithmic) Lognormal as the appropriate resultant distribution by way of the Multiplicative Central Limit Theorem (MCLT). Scientists in each given discipline must agree on such a vista of the phenomenon on hand for this to apply, and such paths haven't been explored yet in the field of Benford's Law. This Multiplicative CLT conjecture for single-issue physical manifestations of the law is all the more appealing given the many extensions, versions, and generalizations of the CLT. Classic CLT states that the sum of various random distributions is Normal in the limit as the number of distributions goes to infinity, and it has three requirements, namely that the various distributions are (I) independent, (II) identically distributed, and (III) have finite variance. There are several generalizations weakening or even eliminating each or some of these three conditions, thus ensuring that the CLT, and by extension the MCLT, has extremely wide applications for real-life physical processes and data. For example, numerous versions are known to generalize the theorem to sums of dependent variables, while others generalize it to non-identical distributions which are bounded in terms of mean and variance, and so forth. One cannot recall another result besides the CLT in mathematical statistics which unites infinitely many processes and distributions into one singular resultant distribution

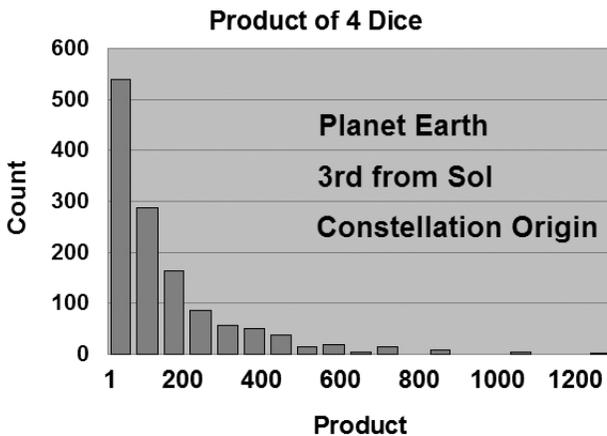
(the Normal); hence the attractiveness of blaming MCLT for such a bewildering large number of manifestations of Benford's Law in the physical world pertaining to totally unrelated and distinct processes all united by their common digital configuration. As was discussed earlier in Chapter 79, taking the logarithm of the product of random processes leads to the sum of the processes, which in turn invokes the CLT in pointing to the Normal as the distribution for that logarithm of the products, and which finally points to the Lognormal as the distribution for the product itself.

In order to illustrate the intimate connection between logarithmically falling density curves having a distinct one-sided tail to the right and multiplication processes, the histogram of one generic example of MCLT process is examined and shown to have exactly such a tail and skewness. In theory MCLT assumes that the number of factors within the product is infinitely large for true convergence of the Lognormal to be obtained, yet, in practical terms, significant convergence is always seen even when only very few multiplication processes are involved. Gambling vices and fraudulent schemes shall serve as the backdrop of the story illustrating skewness. The story refers to a particular casino where four dice are simultaneously thrown and the product of the four faces on the dice serves as the variable determining whether gamblers win or lose. This four-dice variable though is only approximately close to the Lognormal due to insufficient number of multiplications, but it's close enough to serve as a good demonstration. The following four-dice events and their resultant products help to demonstrate the game clearly:

$$\{1, 3, 1, 5\} \rightarrow 15 \quad \{2, 2, 5, 1\} \rightarrow 20 \quad \{3, 4, 5, 5\} \rightarrow 300 \quad \{6, 5, 4, 6\} \rightarrow 720 \\ \{2, 1, 2, 1\} \rightarrow 4 \quad \{1, 6, 5, 1\} \rightarrow 30 \quad \{1, 1, 1, 1\} \rightarrow 1 \quad \{6, 6, 6, 6\} \rightarrow 1296$$

When a large group of wealthy and bored clients walk in, the casino owner entices them to gamble on this new and exciting game of four dice. "Gentlemen, to be more than fair, and to show you our goodwill, the casino wins only if the product is very small, and you lucky fellows win whenever product is either medium or large" he loudly exclaims. When asked how to measure such general and abstract terms he answers: "Let me see here, if all four dice show the small value of 1 then the product is 1; if all four dice show the big value of 6 then the product is 1296 (i.e.  $6^4$ ), and therefore the entire range of possibilities here is [1, 1296]; so let's be more than fair here and you guys win if the product falls in the long interval [301, 1296], while the casino wins only if the product falls within the very short interval [1, 300], and besides, food and drinks are on the house!". Drowned out in the uproar and shouts of excitement was the lone and quiet voice of reason coming from the former chief accountant and fraudster at MF Capital who quickly

calculated and discovered the fraudulent scheme. His simple calculations showed that even the products of middle-dice-values such as three or four are actually quite small within the entire range of possibilities, coming at 81 and 256 (namely  $3^4$  and  $4^4$ ), respectively, and that most people would lose heavily against the casino that night. His insight revealed that most dice products would come well below the 300 cutoff point that the casino owner has intentionally set up. Such is the nature of multiplication that most products fall on the left region of the x-axis where quantities are small, and very few fall on the central or right regions where quantities are medium or big. So much so, that the probability that the product of four dice is small and no more than 300 is 86.9%, while the probability that it's larger than 300 is only 13.1%! The histogram in Fig. 5.2 shows where most products occur, and clearly demonstrates the prominent tail to the right and extreme skewness. The histogram strongly refutes the extremely naïve intuition of a Normal-like symmetric histogram centered on 648.5 [the midpoint of 1 and 1296, namely the average of  $1^4$  and  $6^4$ ]. It also refutes the more sophisticated intuitive naivety of a Normal-like symmetric histogram centered on 150 [the average of  $\{1, 2, 3, 4, 5, 6\}$  taken to the fourth power, namely  $(3.5)^4$ ]. There is nothing particular in having six distinct numbers all equally likely and which increases monotonically by one integer (the dice), and there is nothing unique about having four products as opposed to, say, 17 or 20 products. **The tendency to become skewed to the right is universal in all repeated multiplication processes.** Figs. 1.5 (B) and 1.5 (C) confirm this principle. Figs. 4.58 and 4.59 in Chapters 75 and 76 regarding random and



deterministic flavors also demonstrate and corroborate this strong tendency in the deterministic case of repeated multiplications (of the exponential growth series type). Figure 5.3 depicts the first-digits configuration of the four-dice product, where distribution came out rather close to the logarithmic at  $\{29.2, 16.4, 14.0, 11.4, 4.2, 7.9, 6.8, 3.5, 6.5\}$ , with SSD equals the modest value of 30.5. Superior logarithmic result would have been gotten had the casino owner decided to employ, say, eight dice in a more exciting and more unpredictable game. In a 13-dice game, for example, results should be considered perfectly logarithmic for all practical purposes by any reasonable statistician or data analyst, but not so in the eyes of the stern pure mathematician who would harshly reprimand the statistician for stating such a thing, and who would proclaim that true conformity to Benford's Law can only be found in the limit as the number of dice grows to infinity as predicated by the Multiplicative Central Limit Theorem.

Under the assumption that single-issue physical measures such as earthquake are derived from a product of multiple factors (say in a very simplistic way from the product of numerous tiny tectonics movements), their logarithmic property can be seen as simply a consequence of the Multiplicative Central Limit Theorem. Perhaps a better example of the MCLT conjecture may be given in the population count of a long-established city, where population is expressed as  $P_{\text{TODAY}} = P_0 * F_1 * F_2 * F_3 * F_4$ , and so forth, and  $F_1, F_2, F_3, F_4$ , etc. stand for the yearly random and variable factors expanding or contracting the population, incorporating birth, death, and net migration rates, among other things. Indeed this is identical to the

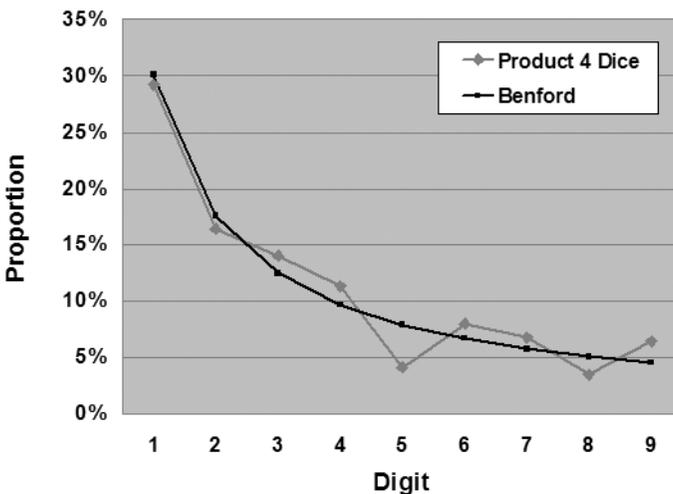


Figure 5.3 First Leading Digits of the Product of Four Dice are Nearly Logarithmic

second model presented by Kenneth Ross and discussed in Chapter 78. It is essential to acknowledge that the principle is by far more general and prevalent than merely population data, and in order to demonstrate this, two additional manifestations of the principle are given. Case I: The final speed of numerous particles of random mass  $M$ , initially at rest, driven to accelerate linearly via random force  $F$ , applied for random length of time  $T$ . In other words, each particle comes with (random) unique  $M$ ,  $F$ , and  $T$  combination. The equations in Physics describing the motion are: (i)  $V_{FINAL} = V_{INITIAL} + A * T$ , or since they start at rest simply  $V_{FINAL} = A * T$ , (ii)  $F = M * A$ . Thus  $V_{FINAL} = (Force * Time) / Mass$ . Monte Carlo computer simulations show strong logarithmic behavior here whenever  $M$ ,  $F$ , and  $T$  are randomly chosen from Uniform distributions of the form  $U(0, b)$ . Case II: The final position of numerous particles of random mass  $M$ , thrown linearly from the same location at random initial speed  $VI$ , under constant decelerating random force  $F$  (say frictional) until each one comes to rest. In other words, each particle comes with (random) unique  $M$ ,  $F$ , and  $VI$  combination. The equations in Physics describing the motion are: (i)  $V_{FINAL} = VI + A * T$ , which leads to  $T_{REST} = VI / A$  as the time it takes to achieve rest, (ii)  $F = M * A$ , (iii) Displacement =  $VI * T - (1/2) * A * T^2$ . Hence Displacement =  $VI * (VI / A) - (1/2) * A * (VI / A)^2$ , namely Displacement =  $(VI)^2 / (2 * (F / M))$ . Monte Carlo simulations again show strong logarithmic behavior here as well whenever  $M$ ,  $F$ , and  $VI$  are randomly chosen from Uniform(0, b). Schematically Case I is of the form Simulation =  $(U1 * U2) / U3$ , while Case II is of the form Simulation =  $(U1 * U1) / (2 * (U2 / U3))$ , and so neither can really apply MCLT, yet digits are almost logarithmic! Conceptually Benford is generally found in nature whenever ‘the randoms are multiplied or divided’. Moreover, even having merely 3, 4, or 5 random multiplied measurements is quite close to Benford, and additional causes such as chain effects and others may further contribute to an even closer convergence as discussed in chapter 95 ‘Hybrid Causes Leading to Logarithmic Convergence’. A decisive demonstration of the validity of the approach in this chapter, and how it can explain an extremely large portion of the physical manifestation of Benford’s Law, is the consideration of simulated results of the product of just two Uniforms, utilizing a distribution form considered ‘anti-logarithmic’ in and of itself.

- Uniform(0, 1)\*Uniform(0, 100) gave{24.1, 18.0, 14.4, 11.9, 9.6, 7.7, 6.5, 4.3, 3.7}
- Uniform(0, 30)\*Uniform(0, 60) gave{28.9, 14.1, 11.3, 10.0, 9.5, 8.3, 7.1, 6.3, 4.6}
- Uniform(0, 33)\*Uniform(0, 70) gave {34.6, 12.4, 10.4, 8.5, 8.0, 7.8, 6.9, 6.0, 5.5}

Simulation results of the product of just two Normal distributions, utilizing a density form that is prevalent in nature, and which is also considered to be 'anti-logarithmic' in and of itself, yield the following results:

Normal(2, 9)\*Normal(5, 13) gave {31.7, 17.6, 11.3, 9.7, 7.8, 6.0, 5.8, 5.4, 4.9}

Normal(4, 7)\*Normal(2, 3) gave {28.7, 18.5, 13.1, 10.2, 7.7, 6.6, 5.9, 4.9, 4.3}

Normal(2, 4)\*Normal(5, 3) gave {29.0, 19.3, 13.3, 10.6, 8.1, 6.4, 4.7, 4.9, 3.7}

Admittedly, the choices of parameters above were somewhat intentional, insuring that the range crosses the origin or that at least it draws plenty from the interval (0, 1) and thus of large order of magnitude, but the general principle here holds nonetheless, since it can be argued that in nature typical Uniforms and Normals are such, often occurring also as tiny fractional quantities near 0 and/or having large OOM, and regardless of the scale in use.

Simulation results of the product of just two Exponential distributions, a density form already somewhat close to Benford in and of itself, yield much superior results:

Exponential(4)\*Exponential(11.0) gave {30.1, 17.6, 12.8, 9.9, 7.3, 7.1, 6.0, 4.7, 4.5}

Exponential(5)\*Exponential(0.07) gave {30.2, 17.5, 12.7, 9.3, 9.4, 7.3, 5.8, 3.8, 3.9}

Exponential(13)\*Exponential(0.2) gave {30.4, 17.1, 12.9, 9.9, 7.5, 6.9, 5.7, 5.3, 4.3}

Hence the only thing the system needs to get nearly logarithmic is nothing but a gentle push forward via other causes or perhaps just an additional product! It should be noted that the nine Monte Carlo simulations above constitute some of the most important results about Benford's Law in this entire book!

The MCLT requires a minimum number of products of distributions in order to obtain something close enough to the Lognormal curve, but often in nature there are typically only 2, 3 or 4 such products, which is not sufficient for MCLT applications. Nonetheless, the mathematics has granted nature convergence in the digital realm, enabling students of BL to let go of MCLT altogether. Examinations of the log curves of the above 9 simulations show that while they are not really Normal-like, they are typically asymmetrical upside-down-U-shaped-like, and the requirements of Related Log Conjecture are almost all met, and thus digits are nearly Benford. On a more profound level, the typical multiplicative form of the equations in physics, chemistry, astronomy, and other disciplines, as well as those of their many applications and results, leads to the manifestation of Benford's Law in the physical world. Newton gave us  $F = M*A$ , not  $F = M+A$ . He gave us  $F_G = G*M_1*M_2/R^2$ , not  $F_G = G + M_1 + M_2-R^2$ , and such is the state of affairs in so many other physical expressions.

## CHAINS AS AN EXPLANATION FOR SINGLE-ISSUE PHYSICAL PHENOMENON

---

The author would like to suggest applying the result of the chain of distributions as another possible explanation for the single-issue physical manifestation in Benford's Law, such as river flow, pulsar frequency, earthquake data, and other variables in chemistry, physics, astronomy, geology, and so forth. This additional conjecture claims that often the physical configuration of the relevant variable corresponds nicely to a chain of distribution, in which the final variable being recorded depends parametrically on other variables. Supporting such an argument is the general observation of the prevalence of causality, dependency, and interconnectedness in the physical world; that often measured variables are themselves parameters for other measured variables.

For example, lengths and widths of rivers depend on average rainfall (being the parameter); and rainfall in turn depends on sunspots, prevailing winds, and geographical location, all serving as parameters of rainfall. Weights of people may depend on overall childhood nutrition, while nutrition in turn may depend on overall economic activity, which in turn depends on economic policy, war and peace, weather-related events such as droughts and flooding, and so forth.

As one specific example, the trajectory of an artillery shell is considered. Here the total destruction of a town in order to save it or to feel the abstract satisfaction of possessing it by firing artillery shells and killing all its civilians is accomplished according to the strict deterministic laws of physics governing projectile motion. Assuming that the height of the glorious cannon itself is negligible, that the muzzle of the cannon is approximately at ground level, and that there is no strong wind affecting the motion, then total horizontal displacement is given by the expression: **Range =  $\sin(2 \cdot \text{Angle}) \cdot (\text{Initial Velocity})^2 / g$** .

The term  $g$  is Earth gravitational constant 9.81, and the Angle is the one between the muzzle and the horizontal ground. In this story, the logarithmic single-issue physical data set in question is the horizontal range of shells of

thousands such cannons on a large battlefield. Random factors such as strong winds, for example, as well as random angular positioning due to lack of targets' visibility render horizontal displacement a random variable having an analytical PDF with Initial Velocity as its singular parameter. Given that all the cannons are of a standard type and having uniform muzzle length, Initial Velocity itself is only a function of the amount of total gunpowder in each shell, and which is not assumed to be standard and fixed but rather random and variable. The patriotic and highly profitable ammunition factory which supplies the cannons with shells benefiting from all the mayhem and slaughter has not automated yet and its operation is heavily labor-intensive. As a result gunpowder is manually shoved into each shell. Therefore, the amount of gunpowder within each shell is randomly distributed approximating the Normal(250, 30), say, in units of grams. This two-sequence chain of distributions, namely the random variable expressing the range **PDF = Range(Initial Velocity(Gunpowder))** is not quite logarithmic due to the insufficient number of sequences in the chain, but it is a great deal similar to it and is at least as near the logarithmic as Stigler's Law is. On the other hand, if gunpowder amount can be assumed to be exponentially distributed (and thus itself quite similar to the logarithmic) then by the extrapolation of the second chain conjecture (to be discussed in detail in Chapter 102) horizontal displacement is nearly logarithmic there and then. Figure 5.4 depicts the outline of the idea of viewing firing distance as a short chain of distributions.

Two competing conjectures have been suggested here that attempt to explain single-issue physical phenomena: (I) repeated multiplication processes, and (II) chains of distributions. Empirical examinations of typical LOG densities of chains of distributions reveal that they are not symmetric, but rather are always slightly skewed to the left, as can be seen in Fig. 4.31. Empirical examinations of typical

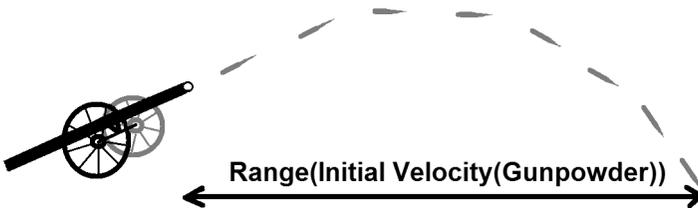


Figure 5.4 Single-Issue Physical Shell Displacement Data as a Chain of Distributions

LOG densities of multiplication processes with large enough number of factors reveal that they are highly symmetric, a fact which is also supported by theory since MCLT predicts having the symmetrical Normal distribution as its LOG density. The fact that almost all single-issue physical data sets that were available for testing empirically by the author came out with near-symmetrical LOG density may lend more credence to the MCLT conjecture. Additional such data sets and their LOG densities must be empirically tested for a more solid conclusion. It is likely that some single-issue physical data sets emanate from multiplication processes while other such data sets emanate from chains of distributions, and that both conjectures are valid.

## BREAKING A ROCK REPEATEDLY INTO SMALL PIECES IS LOGARITHMIC

---

---

Another essential conjecture regarding repeated rock breaking and its resultant strong logarithmic behavior demonstrates once again why the logarithmic is so often found in single-issue physical data sets. The solid confirmations in computer Monte Carlo simulations of logarithmic behavior here, coupled with some mathematical endorsement, render the assertion an actual fact indeed, rather than mere conjecture. Obviously the choice of breaking an actual rock is an arbitrary one, while the generic concept here is the repeated breakup of any quantity into smaller ones over and over again. Once the original whole piece is broken into two pieces, the second stage in the process is the breaking up of the two pieces into four pieces. The third stage consists of the breaking up of these four pieces into eight pieces, and so forth. The essential feature leading to logarithmic behavior here is the random nature in how the pieces are broken. Had a consistent ratio been applied in the breakup such as, say, 50% – 50%, or 30% – 70% in all stages for all the pieces, then no logarithmic convergence is seen whatsoever. At each stage, and for each piece, the determination of the percent breakup is done anew in a random fashion, typically by simulating a realization from the Uniform on  $(0, 1)$  since it can be directly interpreted as percent. In one such actual simulation process employing the Uniform on  $(0, 1)$ , a rock weighing 1000 kilograms was broken randomly and repeatedly in 12 stages into 4096 pieces (namely  $2^{12}$ ). Figure 5.5 shows the particular values gotten in one computer simulation for the first three stages of the process. The table in Fig. 5.6 shows digital configurations for all the 12 stages as well as for the original whole rock composition. Figure 5.7 shows related LOG histogram for all the 4,096 tiny pieces of rocks after the final 12th stage. The wide span on the log axis of roughly 8 units (OOM), and even the more conservative value of roughly 4 OMV units, are strong indications that logarithmic behavior should follow. Digit distribution of the last 12th stage is indeed very close to the

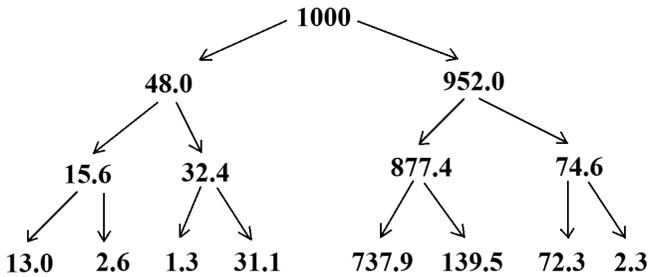


Figure 5.5 Breaking 1000-kilogram Rock into Many Smaller Pieces Randomly

DIGIT:	1	2	3	4	5	6	7	8	9
Original Piece	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2 pieces	0.0	0.0	0.0	50.0	0.0	0.0	0.0	0.0	50.0
4 pieces	25.0	0.0	25.0	0.0	0.0	0.0	25.0	25.0	0.0
8 pieces	37.5	25.0	12.5	0.0	0.0	0.0	25.0	0.0	0.0
16 pieces	43.8	12.5	6.3	6.3	12.5	0.0	6.3	0.0	12.5
32 pieces	34.4	15.6	12.5	9.4	9.4	0.0	3.1	3.1	12.5
64 pieces	26.6	15.6	14.1	15.6	9.4	4.7	4.7	9.4	0.0
128 pieces	29.7	23.4	14.1	8.6	7.0	4.7	4.7	4.7	3.1
256 pieces	33.6	24.6	8.6	8.6	4.7	5.5	4.7	3.5	6.3
512 pieces	32.2	15.6	10.7	9.4	7.2	5.7	5.1	7.0	7.0
1024 pieces	27.7	18.7	11.5	11.2	7.2	7.1	6.3	5.2	5.0
2048 pieces	30.3	18.2	12.0	10.9	7.3	6.3	6.3	4.7	4.0
4096 pieces	30.4	18.1	12.1	9.4	8.3	7.4	5.6	4.6	4.1

Figure 5.6 Digits of the Above 1000-Kilogram Rock Breakdown into Smaller Pieces

logarithmic. Histogram of the data itself regarding those 4096 small rocky pieces (not shown here) is highly skewed with a very long tail to the right. Several other such computer simulations yielded extremely similar result, and so this constitutes a general result. Even better results are gotten with 20 or 30 stages, but these simulations are lengthy and demand high computing power. **Remarkably, this result once again demonstrates how Benford’s Law seems to appear totally out of nowhere!**

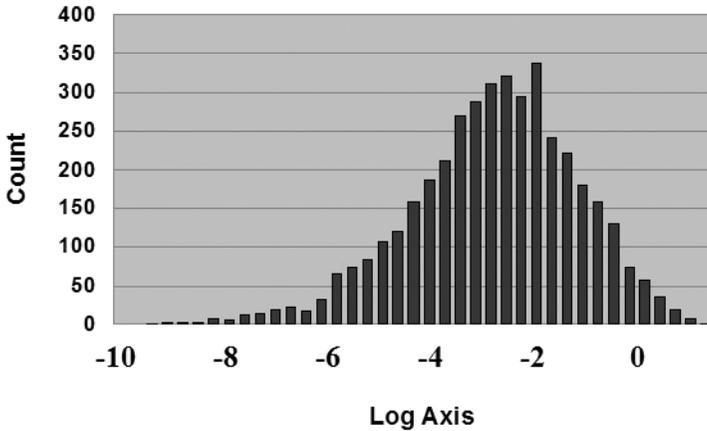


Figure 5.7 Log Histogram of the Above Rock Breakdown — 4096 Pieces of Stage #12

In mathematical terms the process resembles MCLT as it generates the following terms [assuming conveniently that the weight of the original rock is 1 kilogram, and letting  $U_i$  signify a realization from the simulated Uniform on  $(0, 1)$ ]:

$$\begin{aligned}
 &1 \\
 \dots\dots\dots & \\
 &U_1, (1-U_1) \\
 \dots\dots\dots & \\
 &U_2*(U_1), (1-U_2)*U_1, U_3*(1-U_1), (1-U_3)*(1-U_1) \\
 \dots\dots\dots & \\
 &U_4*(U_2*U_1), (1-U_4)*(U_2*U_1), U_5*(1-U_2)*U_1, (1-U_5)*(1-U_2)*U_1, \\
 &U_6*U_3*(1-U_1), (1-U_6)*U_3*(1-U_1), U_7*(1-U_3)*(1-U_1), (1-U_7)*(1-U_3)*(1-U_1) \\
 \dots\dots\dots &
 \end{aligned}$$

And so forth.

Surely this could come under the protective umbrella of MCLT at least approximately since each term is a random partition of  $(0, 1)$ , although some dependencies exist between the terms which complicates the description. In other words, the third stage above is approximately 8 realizations from  $U_I*U_{II}*U_{III}$  albeit with dependencies. For example,  $U_4*(U_2*U_1)$  and  $U_6*U_3*(1-U_1)$  in the third stage are dependent and related via the common value of  $U_1$ . In fact  $U_1$  is present in each of the 8 terms of the third stage, and would always be intertwined in all subsequent stages.

A simulation comparable to the third stage above, with 8 realizations of the product of 3 Uniforms truly in the spirit of MCLT should be as follows:

$$\begin{aligned}
 &U_1*U_2*U_3, & U_4*U_5*U_6, & U_7*U_8*U_9, & U_{10}*U_{11}*U_{12}, \\
 &U_{13}*U_{14}*U_{15}, & U_{16}*U_{17}*U_{18}, & U_{19}*U_{20}*U_{21}, & U_{22}*U_{23}*U_{24}
 \end{aligned}$$

Such proper application of MCLT comes with no dependencies at all, and with many more simulated values from the Uniform, 24 compared with only 7 to be exact, hence the purely MCLT scheme can be termed as 'more random'. Clearly, the process of rock breaking is in need of some more mathematical research and analysis, although the author and his loyal computer are both confident that decisive logarithmic behavior would be part of the conclusion. As mentioned in Chapter 90, students of BL should learn to relax their dependency on that elusive and unnecessary MCLT basis, and rely instead on the more easily available and realistic Related Log Conjecture basis (as strongly suggested by Fig 5.7 here).

Majestic and spectacular supernovas vividly shown on Carl Sagan's *Cosmos* series as well as everyday vicious bomb explosions occurring on this sordid planet have inspired the author in equal measure. Needless to say, many breakdown processes in nature are exemplified by a slow and boring deterioration occurring over enormous stretches of time, and which are logarithmic just the same.

## RANDOM THROW OF BALLS INTO BOXES APPROXIMATING THE LOGARITHMIC

---

---

The opposite act of randomly breaking a rock into small pieces is gluing or attaching small pieces randomly into bigger and bigger chunks of matter (rock formation). One could also consider the random positioning of small particles in space without actually attaching them, like molecules of hot gas moving about inside a container — a classic problem in thermodynamics. When nature creates large things by successive random accumulations of tiny particles or positions numerous molecules in space, results are almost never even and uniform, but are rather skewed in favor of the small and biased against the big. This is a statistical fact that nature cannot overlook, but rather must obey as seen in Thermodynamics and Statistical Physics in general. Such processes of accumulation or positioning yield numerous small quantities and very few big ones, and consequently digit distribution might resemble the logarithmic.

To give a formal statistical structure to a process that attaches and glues things together, a random throw of  $L$  balls into stationary  $X$  boxes is contemplated. All the boxes have the same probability of having a ball landing inside them, namely  $1/X$  per box. In other words, each ball is equally likely to fall into any box, regardless of the number of balls already inside the boxes during each throw. This may differ from random star or planet formation, where bigger chunks of matter are more capable of attracting arriving mass than smaller chunks due to superior gravitational attraction. Specifically, the balls and boxes model assumes that there are no interactions between the balls, neither attraction nor repulsion, and that probability  $1/X$  is unaltered as the boxes get filled.

What happens when balls are thrown into boxes in terms of resultant distribution? In extreme generality results are as follows: if the number of balls is much smaller than the number of boxes ( $L \ll X$ ), then the boxes will end up: (i) mostly empty, (ii) some with only one ball, and (iii) very few rare boxes with two or even three balls. If the number of balls is much larger than the number of boxes ( $L \gg X$ ), then the boxes will end up mostly even, all having about the same

number of balls, namely  $L/X$  on average. In between these two extremes, results show graduation of occurrences being skewed in favor of the small. In fact, even the extreme case of  $L \ll X$  might be characterized as having numerous small ‘quantities’ and very few large ones, assuming one is willing to consider 0 (an empty box) as the smallest of all possible quantitative outcomes of, say,  $\{0, 1, 2, 3, \text{etc.}\}$ .

Let us attempt to aggregate these two extreme cases and some moderate ones in between them to see if anything resembling the logarithmic is obtained in the average distribution. Monte Carlo computer simulations of  $L$  balls thrown into 10 boxes were performed 1,000,000 times (for each value of  $L$ ) and the resultant box proportions recorded. The table in Fig. 5.8 shows resultant box distributions for each  $L$  value in  $\{1, 4, 10, 20, 60, 100\}$ , as well as the average box distribution for the whole set of such  $L$  values (i.e. averaging horizontally row by row). The extreme case of  $L = 1$  is akin to  $L \ll X$  and therefore most boxes are empty. The extreme case of  $L = 100$  is akin to  $L \gg X$  and therefore the biggest portion of the probability occurs with boxes having 8, 9, 10, 11, or 12 balls, namely situations of almost even/uniform box configurations [near the ratio  $100/10$ ]. It must be emphasized that we are not recording digits here (as yet), just counts, from which digit distribution perhaps could be deduced later. The six choices of  $L$  values are arbitrary, yet somewhat similar results for the average are obtained assuming other  $L$  values, unless all are chosen to be too high or too low in relation to the value of 10 (boxes). Upon examining the last column of the average, obviously an empty box is the most probable average value for this particular set of choices of  $L$  and  $X$ , coming at 33.8%, which contrasts sharply with the very low probability of finding a box with numerous balls (e.g. only 3.7% probability of having a box with 8 balls.) Such skewed results, while not exactly logarithmic, are still generally in line with Benford’s Law approximately, especially if one dares to substitute digit 1 for count 0, digit 2 for count 1, digit 3 for count 2, etc., and focuses only on the probabilities of the first 9 ‘quasi-digits’. Such daring an interpretation points to  $\{33.8\%, 17.7\%, 9.5\%, 5.7\%, 4.2\%, 3.9\%, 4.0\%, 3.9\%, 3.7\%\}$ , which adds up to 86.3% only. Adjusting the above proportions by a division by 86.3% we obtain  $\{39.2\%, 20.5\%, 11.0\%, 6.6\%, 4.8\%, 4.5\%, 4.6\%, 4.6\%, 4.3\%\}$ , which does not deviate from the logarithmic a great deal (its 118.6 SSD value is moderate). It must be emphasized again that beyond the particular parameters of  $L \{1, 4, 10, 20, 60, 100\}$  and of  $X \{10\}$ , the resultant average here shown in Fig. 5.8 points to something extremely general, namely numerous empty boxes, some sparsely filled boxes, and very few rare boxes heavily loaded with numerous balls. Admittedly, it

		<b>Number of Balls Thrown</b>						
		1	4	10	20	60	100	Average
		====	====	====	====	====	====	=====
<b>Boxes Having Balls:</b>	<b>0</b>	90.0%	65.6%	34.9%	12.2%	0.2%	0.0%	<b>33.8%</b>
	<b>1</b>	10.0%	29.2%	38.7%	27.0%	1.2%	0.0%	<b>17.7%</b>
	<b>2</b>	0.0%	4.9%	19.3%	28.5%	3.9%	0.2%	<b>9.5%</b>
	<b>3</b>		0.4%	5.8%	19.0%	8.4%	0.6%	<b>5.7%</b>
	<b>4</b>		0.0%	1.1%	9.0%	13.4%	1.6%	<b>4.2%</b>
	<b>5</b>			0.1%	3.2%	16.6%	3.4%	<b>3.9%</b>
	<b>6</b>			0.0%	0.9%	16.9%	6.0%	<b>4.0%</b>
	<b>7</b>				0.2%	14.5%	8.9%	<b>3.9%</b>
	<b>8</b>				0.0%	10.7%	11.5%	<b>3.7%</b>
	<b>9</b>					6.9%	13.0%	<b>3.3%</b>
	<b>10</b>					3.9%	13.2%	<b>2.8%</b>
	<b>11</b>					2.0%	12.0%	<b>2.3%</b>
	<b>12</b>					0.9%	9.9%	<b>1.8%</b>
	<b>13</b>					0.4%	7.4%	<b>1.3%</b>
	<b>14</b>					0.1%	5.1%	<b>0.9%</b>
	<b>15</b>					0.0%	3.3%	<b>0.6%</b>
	<b>16</b>						1.9%	<b>0.3%</b>
	<b>17</b>						1.1%	<b>0.2%</b>
	<b>18</b>						0.0%	<b>0.0%</b>

**Figure 5.8** Random Throw of L Balls into 10 Boxes Resembles the Logarithmic

may seem as if we have been ‘purposely guiding’ the whole scheme and the choices of parameters in the direction towards the logarithmic in some sense, pulling it by the nose, but in spite of appearances the above result stands its ground nonetheless. Nature is frequently chanting ‘small is beautiful’ whenever she forms stars and planets, or arranges her tiny speeding gas particles in space or in the interiors of containers.

In another demonstration of the tendency of random throws to produce skewed and uneven results, the U.S. Federal Reserve Bank is imagined to be attempting to stimulate the economy by printing **\$300,000,000** and actually distributing it randomly to individuals one dollar at a time. It shall be assumed that all **300,000,000** U.S. citizens are eligible and equally likely to get \$1 in the mail, and that such money creation takes place as a long process of repeatedly sending \$1 at a time randomly to citizens without bothering to check whether the

individual has or hasn't gotten previously the same \$1 gift from the Fed. Therefore it is possible that some individuals get multiple checks amounting to multiples of dollars, while others do not receive a single check. Surely this monetary fantasy corresponds exactly to our model of balls (dollars) thrown randomly into boxes (citizens). Simulation of such rare generosity on the part of the Fed shows that 36.7% of the population (110,100,000 people) would receive nothing and stay poor, 36.8% (110,400,000) would receive \$1 only, 18.4% (55,200,000) would receive \$2, while 6.1% (18,360,000) would get \$3. Rarer occurrences are 1.5% of the population (4,560,000) receiving \$4 and 0.3% (906,000) receiving \$5. The luckiest 0.05% of the populations, namely 150,000 people, would receive a total of \$6 dollars in the mail and strike it rich. Still in some even rarer cases more money is gotten. These seven possibilities are highly skewed in favor of low 'digits' (dollars) and resemble the logarithmic philosophically. It must be emphasized that this skewed result strongly depends on the calibration of the parameter input, namely on the total amount of money given by the Fed (in relation to total population). Had the Fed been willing to distribute only one million dollars randomly, then \$0 would almost totally dominate chances, and \$1 occurrences would be relatively very rare. Had the Fed been willing to distribute trillions upon trillions of dollars randomly, inducing an economic boom accompanied by high inflation, then results would be quite egalitarian, and almost every citizen would share equally in the bounty.

[Note: Whenever there is an equality between the number of balls and the number of boxes as in the above Fed scenario (300,000,000), the same box distribution is gotten regardless of the exact number of balls-boxes (so long as it's over 15 approximately). Therefore a small and wealthy municipality with just 25,377 inhabitants imitating the action of the Fed by randomly handing out \$25,377 one dollar at a time would end up giving the sum of dollars of {0, 1, 2, 3, 4, 5, 6} as in {36.7%, 36.8%, 18.4%, 6.1%, 1.5%, 0.3%, 0.05%}, the exact same proportions as in the Fed scenario above. Exceptions are found whenever the number of balls-boxes is small and approximately less than 15, as can be seen in Fig. 5.8 where equality between the balls and the boxes — 10 balls into 10 boxes — yields the box proportions {34.9%, 38.7%, 19.3%, 5.8%, 1.1%, 0.1%, 0.01% }].

Such balls and boxes process and its resultant distribution may serve as a limited model perhaps for U.S. population centers data set which shows strong logarithmic behavior. We contemplate the American continent as being totally empty of humans many millennia ago, and the 'random' arrivals of immigrants from

northeastern Asia walking hungrily and almost frozen over what would become the Bering Strait to Alaska, and from Europe much later, floating over the Atlantic in swaying ships suffering from severe seasickness, and so forth, as balls landing inside boxes. Generic or virtual boxes are imagined as in, say, each 5 square mile patch of area, all collectively covering the entire continent. Yet, since humans tend to congregate and form communities and cities, this is akin to planet and star formation, with attraction and gravitational forces at play, namely an added factor of interaction between the 'balls'.

Oded Kafri, the author of a recent book on entropy, has presented the innovative balls and boxes idea in the context of BL, and claims to have discovered a rigorous mathematical connection between the principles of entropy in physics and Benford's Law for some particular situations where particles are randomly spread throughout certain systems according to the laws of thermodynamics. The author could not independently verify his claim for lack of sufficient knowledge in thermodynamics and in spite of several meetings and discussions with the friendly and patient Kafri. The author was able at least to corroborate approximately partial results via computer simulations, all of which do not seem contradictory to his apparent assertion.

The model presented above is based naturally on letting the balls fall freely into any box, without any bias, prejudice, or 'personal preference' on the part of the falling ball itself in any way. This is what is meant by random and statistical processes, and it guarantees equal chance for any box in successfully calling in a flying ball. Yet the above setup is not the only way to achieve the above-mentioned noble goals of randomness. An alternative approach is to declare all the balls as already residing inside the boxes, and to postulate that all possible box configurations are equally likely. A box configuration is the entire specification of the states of all the boxes, namely the vector of the number of balls within each box. Such a premise does guarantee that no box is unfairly preferred or has some advantage over other boxes in attracting those highly desirable balls. Surprisingly, as it turned out, these two descriptions of randomness are not equal!

The term 'configurational entropy' refers to the assumption that all possible system configurations are equally likely. For example, according to this principle and for the static problem of 15 balls inside 3 boxes, box configuration  $\{5, 5, 5\}$  has the same chance of occurring as, say,  $\{10, 1, 4\}$  or  $\{13, 1, 1\}$ . This premise leads to results that differ from the dynamic process of throwing distinct balls into boxes randomly one by one. Ball throwing leads to unequal probability distribution

of all possible resultant box configurations, contradicting this principle of entropy. Why should a truly random throw of balls yield unequal box configuration? On the face of it, it goes against intuition. The answer is that some configurations are gotten by way of multiple ball throwing scenarios and thus are more easily found, while other configurations are gotten only by way of a singular and unique ball throwing scenario and thus are relatively rare. Kafri chooses thermodynamics' premise of equal configuration probabilities as the basis of his argument that digits here are logarithmic, and that assumption is taken axiomatically. Such a premise implies that even though the system may initially assign identity tags to each ball as well as to each box, distinguishing them, in the final analysis it is not important to know which ball is in which box (in other words, balls lose their identities but boxes retain theirs). Kafri's model focuses only on actual box configuration itself, not on any process that achieves it.

As an example, in the case of 5 balls and 3 boxes ( $L = 5$  and  $X = 3$ ), a final configuration of  $\{L2\}, \{L3, L4, L5\}, \{L1\}$  where the first ball entered the third box, is considered identical to  $\{L1\}, \{L3, L4, L5\}, \{L2\}$  where the first ball entered the first box, and it is considered the same as  $\{L5\}, \{L1, L2, L3\}, \{L4\}$  where the first ball entered the second box, since in all three scenarios above (and in other similar scenarios), at the end of the throwing process, the first box on the left contains 1 ball, the second box in the middle contains 3 balls, the third box on the right contains 1 ball, and that is all that matters. Kafri's model reduces, so to speak, all such scenarios into a singular configuration, namely box configuration  $\{1, 3, 1\}$ , since after all the balls have landed we don't really care about the identities of the balls. Had we gotten, say, the scenario  $\{L1\}, \{L5\}, \{L2, L3, L4\}$ , namely the box configuration  $\{1, 1, 3\}$ , it is considered a distinct configuration though, since boxes retain their identities and the first box on the left is not the same as the third box on the right, or the second box in the middle.

On the other hand, a final configuration of  $\{L1, L2, L3, L4, L5\}, \{\emptyset\}, \{\emptyset\}$  can **only** be gotten by having the first, second, third, fourth, and fifth balls entering the first box on the left, and therefore it is a rarer configuration under the dynamic ball throwing premise where balls retain their identities. This is why in a million repetitions of the game involving the throwing of 5 balls into 3 boxes, box configuration  $\{5, 0, 0\}$  is less likely to occur than box configuration  $\{1, 3, 1\}$ , and this fact (the physical process of ball throwing and its resultant box configurations) is totally independent of whether we humans wish to assign balls identities or do not. Put another way,  $\{5, 0, 0\}$  represents only 1 ball throwing possibility (scenario),

while  $\{1, 3, 1\}$  actually represents 20 ball throwing possibilities (scenarios) and thus the latter comes with 20 times more probability. To summarize: random throws of balls into boxes does not yield equal probability of all possible resultant box configurations, since some box configurations are gotten in many more ways (of ball throwing) than others and thus are more probable. This fact that not all configurations are created equal implies 'increased disorder' philosophically since it points to the fact that extreme concentration of balls into a single box (such as in  $\{5, 0, 0\}$ ) is much rarer than the wide spread of balls among the boxes in a more even fashion.

Hence for a given set of values of  $L$  balls and  $X$  boxes, box distribution depends on the axiomatic direction we take, either (I) dynamical throws of balls into boxes in a truly random fashion, or (II) static with equal probability of all possible box configurations and ignoring balls' identities. Interestingly but not really surprisingly, Benford fares quite well assuming either premise (in extreme generality, considering averaging out a reasonable set of  $L$  values in relations to  $X$ ). The general structure of resultant overall counts is not dramatically different under either premise. Mother Nature still chants her favorite motto 'small is beautiful' regardless of premise. [Yet highly significant differences exist in individual (non-aggregated) distributions between the two premises, as shall be discussed later].

Let's compare the two premises with a concrete example of 5 balls and 3 boxes by observing the different results for each premise. Mathematically, the number of all possible permutations of  $L$  undistinguishable balls in  $X$  distinguishable boxes is  $[(L + X - 1)!] / [L!(X - 1)!]$ . The table in Fig. 5.9 depicts all possible  $[(5 + 3 - 1)!] / [5!(3 - 1)!] = [5040] / [120 * 2] = 21$  box configurations, given 5 undistinguished balls in 3 boxes. Under the assumption that all 21 box configurations are equally likely, and including 0 as one distinct quantity, proportions for  $\{0, 1, 2, 3, 4, 5\}$  are  $\{28.6\%, 23.8\%, 19.0\%, 14.3\%, 9.5\%, 4.8\%\}$ . Kafri suggests as a rule always omitting the 0 possibility altogether, which yields in our case  $\{33.3\%, 26.7\%, 20.0\%, 13.3\%, 6.7\%\}$  as the proportions vector for  $\{1, 2, 3, 4, 5\}$ . Neither proportion is sufficiently near BL Base 6 or Base 7, but overall both distributions are somewhat comparable with the logarithmic. For example, Base 7 distribution in BL is  $\{35.6\%, 20.8\%, 14.8\%, 11.5\%, 9.4\%, 7.9\%\}$ , while Base 6 digit distribution in BL is  $\{38.7\%, 22.6\%, 16.1\%, 12.5\%, 10.2\%\}$ . Kafri further suggests establishing a limit on how many balls can be in a single box, namely an absolute repulsion force between the balls whenever number of balls exceeds a certain value.

{ 5 , 0 , 0 }	0	1	2	3	4	5
{ 0 , 5 , 0 }	0	1	2	3	4	5
{ 0 , 0 , 5 }	0	1	2	3	4	5
{ 4 , 1 , 0 }	0	1	2	3	4	
{ 4 , 0 , 1 }	0	1	2	3	4	
{ 1 , 4 , 0 }	0	1	2	3	4	
{ 0 , 4 , 1 }	0	1	2	3		
{ 1 , 0 , 4 }	0	1	2	3		
{ 0 , 1 , 4 }	0	1	2	3		
{ 3 , 1 , 1 }	0	1	2			
{ 1 , 3 , 1 }	0	1	2			
{ 1 , 1 , 3 }	0	1	2			
{ 3 , 2 , 0 }	0	1				
{ 3 , 0 , 2 }	0	1				
{ 2 , 3 , 0 }	0	1				
{ 0 , 3 , 2 }	0					
{ 2 , 0 , 3 }	0					
{ 0 , 2 , 3 }	0					
{ 2 , 2 , 1 }						
{ 2 , 1 , 2 }						
{ 1 , 2 , 2 }						
	<hr/>					
	18	15	12	9	6	3
	28.6%	23.8%	19.0%	14.3%	9.5%	4.8%

Figure 5.9 All Possible Permutations of 5 Undistinguished Balls in 3 Boxes

[Note: The ordered numbers on the right side of Fig. 5.9 are simply the re-organization of all the values occurring within the boxes seen on the left side. Occurrences are counted and shown just below the line at the bottom. Overall proportions are then calculated and depicted on the last row].

When the premise of actual throws of 5 balls into 3 boxes is taken, results are somewhat different, but still skewed in favor of the small overall. The vector of the proportions for {0, 1, 2, 3, 4, 5} box values is found by simulating one million such complete throws, yielding {13.2%, 32.9%, 32.9%, 16.5%, 4.1%, 0.4%}. As expected, compared to the premise of equal box configurations, the process of actually throwing balls into boxes yields lower probabilities for cases of extreme concentrations such as box configurations {5, 0, 0}, {0, 5, 0}, {0, 0, 5}, since they are rarer, each occurring in only one unique throwing scenario. This explains why chances for 0 and 5 are much lower under the premise of ball throwing than under the premise of equal box configurations. The disparity between the two premises seen here in the particular case of L = 5 and X = 3 about the probabilities of extreme box values 0 and 5 is a general principle, and true for all L balls and X boxes cases, where chances of box value 0 (an empty box) and box value L (a for-

tunate box absorbing all thrown balls) are significantly lower under the ball throwing premise than under the equal configuration premise.

Yet, the particular comparative example above of 5 balls and 3 boxes masks deeper and more significant differences between the two premises in numerous other L and X scenarios. The most dramatic differences occur whenever the number of balls is much larger than the number of boxes ( $L \gg X$ ). In such scenarios, when the premise of ball throwing is taken, the (highly intuitive) expectation is that the boxes will end up mostly even/equal, all having on average about the same number of balls, namely  $L/X$ . But when the premise of equal probabilities of all possible configurations is taken, extreme concentration of all L balls in a single box such as  $\{0, L, 0, 0, 0, \dots\}$  is given serious consideration, an event which is highly unlikely under the ball throwing premise. For example, the scenario of 13 balls ( $L = 13$ ) and 3 boxes ( $X = 3$ ) is one in which there are many more balls than boxes, and where the stark difference between the two premises can be clearly demonstrated.

Possible number of balls in a box:  $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13\}$   
 Throwing balls — % probabilities:  $\{0, 3, 10, 18, 23, 21, 14, 7, 3, 1, 0, 0, 0, 0\}$   
 Equal config. — % probabilities:  $\{13, 12, 11, 10, 9, 9, 8, 7, 6, 5, 4, 3, 2, 1\}$

Here, the difference between the distributions is quite dramatic! Under the ball throwing premise, the expectation is that on average the balls would spread out nicely, filling out the boxes almost evenly. On average, the number of balls in each box should be approximately  $L/X = 13/3 = 4.33$ . Hence we would expect top probabilities to be between 4 and 5 balls, an expectation that is nicely confirmed in the above probability distribution [ $P(4) = 23\%$  and  $P(5) = 21\%$ ]. The chance of throwing all these 13 balls into 1 box, namely configurations of the form  $\{0, 0, 13\}$  is almost zero. Equal configuration premise on the other hand gives such skewed scenarios of the form  $\{13, 0, 0\}$  serious consideration just as it does give to configuration  $\{7, 2, 4\}$  for example.

Another profound difference between the two premises is monotonicity. Equal configurations premise yields strictly monotonically decreasing set of probabilities [consistently skewed in favor of low ball numbers] as can be seen in the distribution above as well as in Fig. 5.9 — assuming that the number of boxes X is more than 2. Without a rigorous mathematical proof to support the above assertion, monotonicity was always empirically found in computer simulated balls and boxes models under equal configuration arrangements, except in  $X = 2$  cases

where the set of probabilities always came out as equal (i.e. uniform). The set of probabilities under the ball throwing premise on the other hand typically (but not always) increases momentarily in the beginning around low ball numbers, then decreases later, as can be seen in the above distribution, as well as in most of the distributions of Fig. 5.8.

Yet, in spite of the profound difference between the two premises, the assertion that Mother Nature always chants her favorite motto ‘small is beautiful’ regardless of premise can be further verified if a table in the spirit of Fig. 5.8 is empirically constructed again for the ball throwing premise as well as for the equal configuration premise, and comparison is made between the two different resultant distributions. Four empirical scenarios are chosen here for analysis, each with a fixed number of 3 boxes, and with varying balls chosen from {1, 3, 5, 13}. The selection of these 4 ball values was made intentionally so as to pick ball values from below as well as from above 3 (the value of the boxes) thus representing a somewhat reasonable and unbiased averaging scheme.

The premise of **equal configurations** yields (% probabilities):

3 Boxes 1 Ball	{67, 33, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
3 Boxes 3 Balls	{40, 30, 20, 10, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
3 Boxes 5 Balls	{29, 24, 19, 14, 10, 5, 0, 0, 0, 0, 0, 0, 0, 0}
3 Boxes 13 Balls	{13, 12, 11, 10, 9, 9, 8, 7, 6, 5, 4, 3, 2, 1}
<b>Average:</b>	<b>{37, 25, 13, 9, 5, 3, 2, 2, 1, 1, 1, 1, 0, 0}</b>

The premise of **ball throwing** yields (% probabilities):

3 Boxes 1 Ball	{67, 33, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
3 Boxes 3 Balls	{30, 45, 22, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
3 Boxes 5 Balls	{13, 33, 33, 16, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0}
3 Boxes 13 Balls	{0, 3, 10, 18, 23, 21, 14, 7, 3, 1, 0, 0, 0, 0}
<b>Average:</b>	<b>{28, 29, 16, 10, 7, 5, 3, 2, 1, 0, 0, 0, 0, 0}</b>

Hence, while results of the average for the two premises are different somewhat, both premises lead to the same overall structure of distribution where the small is favored and the big is rather rare, and this fact is general, true for other similarly constructed balls and boxes schemes with comparable parameters.

The generic scheme in this chapter was always constructed by fixing the number of boxes while varying the number of balls, followed by the averaging out of the probabilities. It is hard to judge whether or not such a scheme is akin to planet and star formation or other processes found in nature. It is certainly possible that

an inverse scheme may be more fitting in some particular physical processes, namely the scheme that varies the number of boxes and fixes the number of balls. Yet, such inverted schemes yield almost the same overall structure in results shown in this chapter. As for one concrete example applying the ball throwing premise, the number of balls is fixed at 9, while the number of boxes varies as in the set {3, 6, 9, 17, 33}. The values of the boxes were intentionally chosen around the central value of 9 (the fixed number of balls) so as to create an unbiased scheme with a reasonable and fair spread in the number of boxes compared with the fixed number of balls. These five schemes of 9-ball throws yield (% probabilities):

9 Balls 3 boxes	{ 3, 12, 23, 27, 21, 10, 3, 1, 0, 0 }
9 Balls 6 boxes	{19, 35, 28, 13, 4, 1, 0, 0, 0, 0 }
9 Balls 9 boxes	{35, 39, 20, 6, 1, 0, 0, 0, 0, 0 }
9 Balls 17 boxes	{58, 33, 8, 1, 0, 0, 0, 0, 0, 0 }
9 Balls 33 boxes	{76, 21, 3, 0, 0, 0, 0, 0, 0, 0 }
<b>Average:</b>	<b>{38, 28, 16, 9, 5, 2, 1, 0, 0, 0 }</b>

The average of the five schemes is strictly monotonically decreasing [last values were rounded as zeroes], and we still obtain the same overall structure in resultant distribution, namely numerous empty boxes, some sparsely filled boxes, and very few rare boxes heavily loaded with numerous balls. This last result together with all the previous results arithmetically imply that balls and boxes schemes where both L and X parameters vary concurrently yield the same distribution structure. Therefore natural processes fitting simultaneously varying balls and varying boxes schemes necessarily come with a quantitative distribution of the same structure where the small is numerous and the big is scarce.

On the face of it, Kafri's choice might seem arbitrary, as it discriminates against balls' identities while favoring boxes' identities. Yet, this is actually quite natural in thermodynamics where molecules serve as balls while spatial and volume segments serve as boxes in the model. When by pure chance hot gas is suddenly and spontaneously pressing against the valve on the upper right side of the metal container to the breaking point, we do not care 'which' molecules are involved, only that the space around the valve (that 'box') has a very high concentration of molecules and therefore pressure is particularly high there. Had this spontaneous pressure been applied to the lower left side (another 'box') against the solid strong metal, it would constitute a distinct scenario, and this is why positions/boxes have identities. In other words, it is quite natural in thermodynamics to focus solely on resultant configurations, not on the particular molecular composition. Even if we could identify the molecules by way

of some tiny unique signature, thermodynamics would show no interest at all in pursuing such a path. Pressure is purely a function of (i) average speed of molecules, (ii) average weight of molecules, (iii) number of molecules moving in a given volume, and in the context of thermodynamics molecules' identities are totally irrelevant. When the owner of a heavy truck with an engine out of order hires 8 men to push it up the hill, and then hires another group of 8 men the next day to get it across town to the mechanic shop, the personalities, names, Social Security Numbers, etc. of the working men are unimportant to him as he focuses solely on the task of securing and moving his precious truck. Yet, they are distinguishable and unique in other contexts, and certainly to themselves! The concepts of replacement and substitution should not be confused with the concept of lack of identity.

In addition, there is a profound scientific reason for discriminating against balls' identities in the context of thermodynamics. It is certainly possible to sort through a sack of rice and give an identity to each and every grain by weight, composition, shape, and so forth. The same can be done for snowflakes. Yet, current technology cannot do that for 'identical' atoms of the same isotope types, and theoretical physicists would be up in arms and quite upset should anyone attempt such an atom-identifying scheme. Are two oxygen elements of the most common isotope  $^{16}\text{O}$  really totally indistinguishable? Could it be that there is still some tiny variability in, say, atomic weight and that no two atoms weigh exactly equally (enabling us to use weight as an 'atomic signature')? Can two water molecules of identical isotope types (being a bit more complex) be distinguishable by the exact angle measure between the two hydrogen atoms? Or is that angle fixed exactly at 104.45 degrees in the ground state for any water molecules? The whole science of Quantum Mechanics is predicated on whole exact quantities and qualities and it precludes any possibility of assigning identities to elements (and certainly much less so to those 'smaller' and more basic sub-atomic 'particles').

Intuition strongly suggests that the model of actual ball throwing premise should yield identical results to another physical model where balls start out residing in the boxes having any initial configuration whatsoever, followed by the subsequent process of moving one ball at a time, from one box to another box, randomly and indefinitely. The ball-removal aspect of such random exchange process ('out') is not based on (equally chanced) box selection, but rather on (equally chanced) ball selection, with the possibility of removing ball  $L_i$  from box  $X_j$  and returning it to the same  $X_j$  box from which it came from. The ball-return aspect ('in') is based on (equally chanced) box selection. Computer simulations decisively confirm this compelling intuition, yielding practically identical results for both physical processes.

As a final note, regarding Fig. 5.8 of  $L$  balls into  $X$  boxes, it might be suggested to put the whole model on a chain of distribution basis in order to obtain better results. A superior model would randomly pick the number of  $L$  balls from, say, the discrete Uniform on  $\{1, 2, 3, \dots, X, \dots, (2X - 3), (2X - 2), (2X - 1)\}$  which is nicely centered around the value of  $X$  boxes. For example, for  $X = 10$  boxes,  $L$  is randomly chosen from the discrete Uniform on  $\{1, 2, 3, \dots, 17, 18, 19\}$ , which leads to the box distribution [beginning with an empty box having 0 balls] of:  $\{61\%, 27\%, 9\%, 2\%, 1\%, 0\%, 0\%, 0\%, \text{etc. to } 19\}$ . If instead  $L$  is randomly chosen from the discrete Uniform  $\{1, 2, 3, \dots, 98, 99, 100\}$ , having the same  $X = 10$  number of boxes, and which is much more in the spirit of the setup of the table in Fig. 5.8, resultant box distribution is [beginning with an empty box of 0 balls]:  $\{26.0\%, 19.8\%, 14.3\%, 10.9\%, 8.3\%, 6.3\%, 4.7\%, 3.4\%, 2.4\%, 1.6\%, 1.0\%, 0.6\%, 0.4\%, 0.2\%, 0.1\%, 0.1\%, 0.0\%, 0.0\%, 0.0\%, \text{etc. to } 100\}$ , and upon deleting the empty 0 ball occurrences of 26.0% [and dividing the rest by the total of 74%], we finally get the quantitative distribution [beginning with the non-empty case of 1 ball]:  $\{26.8\%, 19.3\%, 14.7\%, 11.2\%, 8.5\%, 6.3\%, 4.6\%, 3.2\%, 2.1\%, 1.4\%, 0.8\%, 0.5\%, 0.3\%, 0.1\%, 0.1\%, 0.0\%, 0.0\%, 0.0\%, \text{etc. to } 100\}$ . Converting the averaging process built into the table of Fig. 5.8 into the discrete Uniform distribution reminds one of the chains of distributions, and that such representation and interpretation of averaging scheme has led to additional results and better insight. Unfortunately not much is gained here under such interpretation. Even if one simplistically attempts to imitate the second chain conjecture and replaces the Uniform here with something having its digits resembling the logarithmic, not much is gained. For example, for number of boxes fixed at  $X = 13$ , and number of balls varying randomly as in the (integral values of) exponential 30% growth from base 1 of the set  $\{1, 1, 2, 2, 3, 4, 5, 6, 8, 11, 14, 18, 23, 30, 39, 51, 67, 87\}$ , results are still disappointing at  $\{45.4\%, 20.6\%, 10.6\%, 7.0\%, 5.0\%, 3.7\%, 2.7\%, 1.9\%, 1.3\%, 0.8\%, 0.5\%, 0.2\%, 0.1\%, 0.1\%, 0.1\%, 0.0\%, 0.0\%, \text{etc. to } 87\}$  where 45.4% stands for the probability of an empty box with 0 balls. Upon deleting the empty box occurrences of 45.4% [and dividing the rest of the values by the total of 54.6%], we get the following results: [beginning with the non-empty case of 1 ball]  $\{37.7\%, 19.5\%, 12.7\%, 9.2\%, 6.8\%, 5.0\%, 3.5\%, 2.4\%, 1.5\%, 0.9\%, 0.5\%, 0.2\%, 0.1\%, 0.1\%, 0.0\%, 0.0\%, \text{etc. to } 87\}$ . As expected, in all of these trials and twists the general principle of numerous small quantities and few big ones still holds firm!

## LOGARITHMIC MODEL FOR PLANET AND STAR FORMATIONS

---

---

Intuition demands that if the logarithmic was directly found in **destructions** (rock breaking) then it should also be directly found in **constructions** (balls or pieces of mass forming). Kafri's model of throwing balls into boxes is a well-structured process that leads only to logarithmic-like behavior in the approximate, and only under certain averaging schemes. Each thrown ball is certain to land inside some box. If after plenty of ball throwing there are no empty boxes, then subsequent balls are certain to get glued to other previously thrown balls with 100% probability. The other extreme is a model without any structure for boxes whatsoever, where balls are thrown in space and do not get connected at all [0% probability of connectivity]; they simply float out there alone and we are left with a collection of values of 1's as the dataset. In between these two extreme scenarios [100% chance of entering into a box, and 100% chance of staying alone and unattached], a process without any boxes is envisioned where each thrown ball or piece of mass has  $P$  probability of getting attached to other standing balls or masses, and  $(1 - P)$  probability of continuing to float out there in space alone not connecting to existing balls or masses. Such a process is more appropriate for star and planet formations than the well-structured balls into boxes model, since mass arriving into, say, a nebula does not necessarily get attached to existing mass, and this aspect of the process is probabilistic rather than deterministic. In addition, it is necessary to postulate a random quantity of arriving mass, say from [Kilogram]\*[Uniform(0, 1)], as opposed to the exact unity of the balls and boxes model [where arriving ball contributes the exact whole value of 1]. A random value of arriving quantity is a more realistic model in star and planet formation where arriving mass is of variable and random weight. The third (implicit) assumption for this model is that even though masses (balls) do interact in real life via gravitational forces, effectively that interaction is not so much a function of existing mass [as predicated by the gravitational formula] but rather is mostly a function of the (random) trajectory of arriving mass which may be aiming or not aiming at

existing mass. Since gravitational force is inversely proportional to the distance, its sphere of influence is really mostly in the immediate vicinity of the object, hence what really matters here is the trajectory of the incoming mass and how close it approaches existing mass. In conclusion: arriving mass gets attached to existing mass as in the discrete Uniform and is equally likely to choose any existing mass without preferring heavier ones [i.e. equal ability of all existing pieces of mass in attracting newly arriving mass — regardless of their weight].

Indeed, computer simulations show that such a process leads directly to the logarithmic distribution without any need to average things out in any particular way, as was done for the balls and boxes model. The following simulation results pertain to pieces of mass randomly chosen from the Uniform on  $(0, 1)$ , and thrown with  $P$  probability of attaching itself to something. The first piece of mass is just being put out there without the possibility of gluing it to anything. It is only from the second piece of mass onwards that the process starts in earnest. The condition  $P > 0.50$  is assumed, and so since most pieces are fused together, the original number of created and thrown pieces are reduced in the process, and this is indicated as [(original # of created pieces)  $\rightarrow$  (final # of pieces)]. The table in Fig. 5.10 depicts the first- and second-digits distributions results for six distinct simulations having a variety of  $P$  values. Since  $P$  represents the proportion of the original number of created pieces that get fused,  $(1 - P)$  represents the proportion of them that are set apart creating new entities, thus (Original # of Pieces)\* $(1 - P)$  should equal approximately the final number of pieces in the system at the end of computer simulations. For example, in the first row  $P = 0.97$ , (Original # of Pieces) = 90000, hence we expect about  $(90000)*(1 - 0.97) = 2700$  to exist at the end, an estimated value which is quite near the actual value of 2682 observed in the simulation. The last column (in light gray color) is the misguided simulation assuming a very low  $P$  value of 0.09 for the chance of pieces getting glued. If that probability is made to be so low, most of the pieces created fly out there without getting attached to anything; almost no star or planet is being born; almost no reduction in the number of pieces is observed, and the digit distribution of the final system is almost identical to that of the generating process, namely the uniform digital configuration of the Uniform $(0, 1)$  distribution.

**Remarkably, this result once again demonstrates how Benford's Law seems to appear totally out of nowhere!** The result seen in Fig. 5.10 is parameter-free and independent of any input, so long as  $P$  is set high enough! Digit

P = 0.97 [ 90000 → 2682 ]	31.5	18.0	13.2	9.1	7.0	6.1	6.1	4.7	4.4	
P = 0.93 [ 40000 → 2669 ]	30.6	16.7	11.5	9.0	7.4	6.4	7.2	6.0	5.1	
P = 0.90 [ 40000 → 3904 ]	29.6	15.0	13.1	10.6	7.3	7.3	5.6	6.7	4.9	
P = 0.75 [ 15000 → 3882 ]	30.0	18.8	13.1	8.0	7.0	5.9	5.7	5.8	5.6	
P = 0.67 [ 50000 → 16531 ]	31.3	17.7	11.2	7.7	6.7	6.2	6.3	6.3	6.6	
P = 0.09 [ 13000 → 11823 ]	14.6	10.4	10.5	10.8	10.4	10.4	10.7	10.9	11.3	
<b>Benford's Law 1st Digits:</b>	<b>30.1</b>	<b>17.6</b>	<b>12.5</b>	<b>9.7</b>	<b>7.9</b>	<b>6.7</b>	<b>5.8</b>	<b>5.1</b>	<b>4.6</b>	
P = 0.97 [ 90000 → 2682 ]	11.5	11.7	11.1	10.4	9.2	9.1	9.4	9.8	8.7	9.2
P = 0.93 [ 40000 → 2669 ]	11.5	11.7	10.6	11.5	9.8	9.9	7.9	8.8	8.7	9.6
P = 0.90 [ 40000 → 3904 ]	12.8	12.4	10.5	10.5	9.3	10.1	8.3	9.7	8.1	8.4
P = 0.75 [ 15000 → 3882 ]	11.2	11.1	10.5	11.5	9.5	9.0	9.3	10.2	8.9	8.8
P = 0.67 [ 50000 → 16531 ]	11.5	10.9	10.9	10.2	10.0	10.1	9.5	8.9	9.2	8.7
P = 0.09 [ 13000 → 11823 ]	10.5	10.1	10.7	10.2	9.7	9.7	9.8	10.1	9.8	9.3
<b>Benford's Law 2nd Digits:</b>	<b>12.0</b>	<b>11.4</b>	<b>10.9</b>	<b>10.4</b>	<b>10.0</b>	<b>9.7</b>	<b>9.3</b>	<b>9.0</b>	<b>8.8</b>	<b>8.5</b>

Figure 5.10 Digital Results of Monte Carlo Simulations of Star and Planet Formations

distribution converges rapidly to its expected logarithmic configuration without waiting for more pieces to arrive, hence beyond the first few hundred or so throws, the number of pieces is not a factor at all in the eventual digital configuration that develops. To emphasize this point once again: converging digit distribution is totally independent of how many pieces are created and thrown! Surely, the more pieces we throw, the more pieces exist, but the logarithmic property is a constant that becomes steady after the first few hundred or thousand throws. The system obtains the logarithmic property and it is not dependent on anything, including the data distribution that originates the mass which, in our case, is the Uniform(0, 1). Had other mass-generating distributions been employed, the same result would have been gotten! Indeed, if the exponential or the Lognormal distributions were to be employed, stronger (or rather, faster) logarithmic result would be observed. The motivation for the choice of the Uniform is to demonstrate the powerful tendency towards the logarithmic in spite of the fact that generated values are decidedly non-logarithmic, since the Uniform may be considered to be 'anti-logarithmic' due to its digital equality property. The star and planet formation model of this chapter is superior to the balls and boxes model in six ways: (1) results are extremely close to the logarithmic by far more so than any balls and boxes scheme, (2) there is no need to substitute digits for counts, (3) there is no need to arbitrarily cut off the long vector of count proportions at the ninth or

tenth place, (4) there is no need to debate and decide on what should be done with an empty box of 0 value, (5) the model is parameter-less since there is no need to calibrate any numbers such as L balls and X boxes values, nor anything else, (6) and most importantly: the logarithmic is obtained directly; there is no need to average out distinct digits/counts distributions pertaining to distinct L and X scenarios.

Following the process closely, at each stage, as pieces gradually build up their masses, some of the mystery can be taken out of it. As new mass arrives at the scene and simulated value for P indicates that it should get attached to an older existing piece, all pieces are equally likely to add to themselves this available arriving mass. Once in a while simulated value for P indicates that arriving mass should stand alone as a new piece. Thus all existing pieces at any stage come with a strict hierarchy of 'age' [age counted in 'stage-years']. No two pieces exist having the same age [i.e. there are no twins]. Relatively older pieces gain on average many times more arriving mass than younger ones, while very recent pieces [infants] have had almost no chance in attracting arriving pieces since they have been in existence for only a very short time. The net effect of this dichotomy between the new and the old is differentiated mass values, hence the small is definitely more beautiful than the big, and all this conceptually explains why the logarithmic emerges here.

Besides Monte Carlo computer simulations, real-life empirical confirmation of sorts for this model can be found in data regarding exoplanets' mass which came out very close to the logarithmic (Chapter 10, Fig. 1.21). The author did not have a chance yet to perform empirical tests on star mass data, although results should almost certainly show strong logarithmic behavior there as well.

## HYBRID CAUSES LEADING TO LOGARITHMIC CONVERGENCE

---

Various causes were shown to lead to Benford behavior in and of themselves, such as (1) mixing totally unrelated numbers from a variety of sources (Hill), (2) having data starting from Lower Bounds values such as 0 or 1 while Upper Bounds vary gradually upwards (averaging schemes), (3) chaining parameters to other distributions, (4) exponentially growing quantities, (5) multiplying the same distribution or data type by itself numerous times (MCLT), (6) selecting combinations from a long list of numbers (Random Linear Combinations), (7) the repeated random breakup of a quantity into numerous smaller ones (rock breaking), (8) the random throw of balls into boxes, as well as other causes.

It is suggested that at times the logarithmic is found in real-life data sets by way of the combined effects of several incomplete such causes, with each cause individually being too diluted or weak to lead to full logarithmic convergence by itself. Such a hybrid of causes could lead to a strong convergence in a cumulative way.

As an example, the following short chain is considered:

**Normal**(Uniform<sub>1</sub>(0, 1)\*Uniform<sub>2</sub>(0, 1), Mixed Collection of 1000 numbers)

This process represents three distinct causes: (1) one weak manifestation of a chain of distributions, (2) a very limited multiplicative process (MCLT), and (3) a small random collection from a variety of data sources (Hill), all of which are combined to form a singular data set having much stronger logarithmic behavior than each one of its three constituent causes alone. It should be noted that the Normal distribution fully converges to the logarithmic only whenever both parameters — the mean as well as the standard deviation — are chained (in an infinite manner or to some logarithmic distributions) as shall be discussed in a later chapter on chainable parameters. A related consequence (the extrapolated second conjecture) is that the closer to the logarithmic the two parametrical distributions are, the closer is the Normal to the logarithmic as well (i.e. a direct relationship

between logarithmic-ness of the chaining distributions, and the logarithmic-ness of the chained distribution).

The mean parameter of this Normal distribution is obtained from 1000 products, each being a realization of simulation from the Uniform on (0, 1) times another such realization of simulation from the Uniform on (0, 1). Since the product  $U_1 * U_2$  involves only two elements being multiplied, it does not represent a truly multiplicative process applicable to MCLT in the context of BL, yet it is somewhat not too far from being logarithmic. The set of 1000 mixed numbers is not fully logarithmic because of two reasons: (I) collection was not done perfectly according to Hill's model in the sense of truly varying the source constantly, as many values were obtained from identical sources, (II) the size of 1000 is a bit small in a statistical sense. Several simulations gave nearly identical results of the first digit. Here are the digital results of one such simulation:

$U_1(0, 1)$ : {11.0%, 10.0%, 11.7%, 13.1%, 10.9%, 10.9%, 10.1%, 12.4%, 9.9%}  
 $U_2(0, 1)$ : {11.6%, 13.1%, 11.7%, 10.4%, 11.2%, 10.2%, 9.8%, 10.3%, 11.7%}  
 $U_1 * U_2$ : {24.5%, 20.5%, 14.3%, 9.5%, 9.4%, 7.2%, 5.5%, 5.3%, 3.8%}  
Mixed Coll.: {25.3%, 15.3%, 14.3%, 10.0%, 8.8%, 7.8%, 7.3%, 6.2%, 5.0%}  
 $N(U_1 * U_2, M)$ : {30.0%, 17.3%, 12.1%, 10.2%, 8.0%, 7.1%, 5.1%, 5.1%, 5.1%}

The close convergence to the logarithmic of  $N(U_1 * U_2, \text{Mix})$  is due to the fact that both parameters are tied to distributions that are in and of themselves somewhat close to the logarithmic. This is but one demonstration of the possibility of hybrid causes leading to resultant logarithmic behavior in real-life data sets.

## MILD DEVIATIONS SEEN IN SMALL SAMPLES OF LOGARITHMIC DATA SETS

---

---

In order to demonstrate the tenacity and prevalence of the logarithmic distribution in real-life data, eight small data sets (of the logarithmic type) having only about 30 points each shall be examined. This illustration aids in informally assessing typical magnitudes of deviation from the logarithmic that occur in data sets due to their small size. The table in Fig. 5.11 contains digital results from very small data sets having only 30 to 32 points, drawn randomly from some large logarithmic population data or generated via processes that lead to the logarithmic. Samples of 30 values each were randomly selected from data sets mentioned earlier regarding time between earthquakes, U.S. population centers, and the mixture of unrelated 34,269 numbers collected in the spirit of Hill's model. Simulations of the Lognormal with location 5 and shape 1, the rather short chain  $\text{Uniform}(0, \text{Uniform}(0, \text{Uniform}(0, 7)))$ , as well as seven-dice products as in MCLT, were all limited to 30 realizations. In addition computer simulations of 30 chemical compounds were realized as in Chapter 88, selecting randomly from the first 35 elements in the Period Table up to Bromine as in the discrete Uniform  $\{1 \text{ to } 35\}$ , with frequency as in the discrete Uniform  $\{1 \text{ to } 5\}$ , having two or three elements decided upon with the flip of a coin, representing one Random Linear Combination case. Lastly, a rock weighing 25 kilograms was randomly broken into 32 pieces in five stages via the  $\text{Uniform}(0, 1)$ . Not surprisingly, digital results for these eight small data sets do not deviate much from the logarithmic, since they are all derived from eight population datasets and processes that are in and of themselves extremely close to the logarithmic.

Remarkably, the digital average of the eight small data sets (shown in Fig. 5.11) came out nearly logarithmic with a very low SSD value of 9.1. These values were calculated as the simple averages column by column in Fig. 5.11, and are extremely close to the digital proportions calculated for all 242 data points combined as one single data set. The strong logarithmic behavior of this average can be explained in two ways: (I) either as in the spirit of Hill's model of distribution of many

Data Set	1	2	3	4	5	6	7	8	9	SSD
30 US Population Centers	33.3	13.3	13.3	6.7	10.0	6.7	6.7	6.7	3.3	47.6
Time between 31 Earthquakes	23.3	16.7	10.0	10.0	10.0	3.3	10.0	13.3	3.3	155.4
Breaking a Rock 5 Times	46.9	12.5	6.3	9.4	3.1	6.3	3.1	9.4	3.1	397.1
30 Products of 7-Dice Throws	23.3	23.3	13.3	13.3	6.7	6.7	3.3	3.3	6.7	107.8
30 Simulated Chemicals	33.3	26.7	13.3	10.0	6.7	3.3	3.3	0.0	3.3	139.9
Lognormal Loc = 5 Shape = 1	23.3	13.3	13.3	20.0	3.3	0.0	13.3	10.0	3.3	319.1
30 Numbers from Mixed Sources	26.7	20.0	20.0	6.7	10.0	3.3	6.7	3.3	3.3	104.1
Chain of Dist: U(0, U(0, U(0, 7)))	36.7	13.3	13.3	13.3	10.0	6.7	3.3	0.0	3.3	113.5
=====	=====	=====	=====	=====	=====	=====	=====	=====	=====	=====
Average Digits of 8 Data Sets	30.9	17.4	12.9	11.2	7.5	4.5	6.2	5.8	3.7	9.1
Benford's Law, LOG(1+1/d)	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6	0.0

**Figure 5.11** Digital Deviations Are Not Extreme Even for Very Small Data Sets

distributions, since those 242 values are derived from eight very different sources, or (II) simply representing an overall much larger sample size of 242 data points (30\*7 plus 32 broken rocks) and thus easily inheriting the logarithmic property of the larger original data sets and processes. It is reasonable to conjecture that Hill's model converges faster to the logarithmic when distributions themselves are logarithmic to begin with (or just close to it), although crucially Hill's liberal and inclusive model is based on all types of densities, logarithmic and otherwise, including the Normal, the Uniform, as well as 'anti-logarithmic' ones where digital order is inverted in favor of high digits [so long as they are defined on the positive x-axis exclusively].

## THE REMARKABLE VERSATILITY OF BENFORD'S LAW

---

Let us end this section by noting the remarkable versatility and prevalence of Benford's Law in the natural world as well as in the realm of its abstract manifestations:

Break it up — it's Benford.

Glue it together — it's Benford.

Spin it around (rotation frequency of pulsars) — it's Benford.

Pick it up blindly and indiscriminately (Hill's model) — it's Benford.

Multiply it — it's Benford.

Divide it — it's Benford.

Chain distributions up — it's Benford.

Combine many data sets together — it's Benford.

Spread it around (populations and such) — it's Benford.

Shake it up (earthquakes) — it's Benford.

This is a very curious state of affairs; seemingly totally distinct and unrelated processes lead to the same numerical results:  $\text{LOG}(1+1/d)!$  It is quite astonishing! Imagine seven chefs — French, Thai, Russian, Chinese, Japanese, African, and Indian — on different continents, using totally different ingredients, cooking styles, pot sizes, and flame sizes. Yet all seven come up with seven dishes equal in taste, consistency, flavor, smell and colors! In the same sense it appears far-fetched that totally different physical processes and abstract models yield the same exact digit distribution for all the nine digits! There must be some fundamental commonality and correspondence between all of them!

The ultimate quest in Benford's Law is to somehow unite all these diverse physical processes and operations into a singular concept and show that that fundamental concept is logarithmic. Such a goal may prove elusive, though, if nature and/or the involved mathematics really tell the story of several distinct causes all

leading to the same logarithmic conclusion. Surely Mother Nature would be immensely upset upon learning that someone is attempting to simplify or belittle her complex and versatile digital behavior. Mathematicians wishing to take on this challenge should be warned — she is quite malicious and vindictive when provoked!

## **Section 6**

### **TOPICS IN BENFORD'S LAW**

**This page intentionally left blank**

## SINGULARITIES IN EXPONENTIAL GROWTH SERIES

---

Exponential growth series are expressed as  $\{\mathbf{B}, \mathbf{Bf}, \mathbf{Bf}^2, \mathbf{Bf}^3, \dots, \mathbf{Bf}^N\}$ , where  $\mathbf{B}$  is the base (initial) value,  $\mathbf{P}$  is the percent growth, and  $\mathbf{f}$  is the constant multiplicative factor relating to the growth rate as in  $\mathbf{f} = (\mathbf{1} + \mathbf{P}/100)$ . Throughout this chapter we shall operate under the assumption of our decimal base 10 number system, to be generalized for other bases and further discussed in Section 7. The related (common) log series is simply:

$$\{\text{LOG}_{10}(\mathbf{B}), \text{LOG}_{10}(\mathbf{B}) + \text{LOG}_{10}(\mathbf{f}), \text{LOG}_{10}(\mathbf{B}) + 2*\text{LOG}_{10}(\mathbf{f}), \\ \text{LOG}_{10}(\mathbf{B}) + 3*\text{LOG}_{10}(\mathbf{f}), \dots, \text{LOG}_{10}(\mathbf{B}) + N*\text{LOG}_{10}(\mathbf{f})\}.$$

Due to their discrete nature, (finite) exponential growth series are never exactly logarithmic, only in the limit possibly. Discreteness in general causes deviations from the logarithmic, and could possibly be remedied by considering infinitely many sequences and the limiting case. The understanding gained earlier (Proposition VI) about the nature of  $k/x$  distribution and its intimate connection to deterministic multiplication processes hints at the necessity to stand between two IPOT values in the case of exponential growth series as well (or more generally, simply having an integral exponent difference), its discrete nature notwithstanding. This necessity is exactly what is empirically found when a large variety of exponential growth series are computer-simulated, providing a strong confirmation for the theoretical framework developed earlier. Two factors facilitate and determine logarithmic behavior for exponential growth series: (1) the length of the series, which normally must be sufficiently long and contain enough sequences for a close logarithmic fit, (2) the value of exponent difference between the first and the last elements (*or equivalently, the log difference in the two extreme values*), which should be ideally as close to an integer as possible for a good logarithmic fit. In general, low growth series that adhere closely enough to the above two requirements are close enough to the logarithmic. For very high growth series, where the digital cycle is very short (i.e. passing IPOT points frequently), having an integral exponent difference is not sufficient at all or even necessary; rather it is necessary

to have sufficient number of sequences, and only for such long series is the logarithmic approximately observed.

It is certainly proper to think of  $\text{LOG}(f)$ , namely  $\text{LOG}(1+P/100)$ , as being a fraction, as it is so in any case up to 900% growth of  $P$ , else the integral part can be ignored as far as digital configuration is concerned. For example, for 2% growth,  $\text{LOG}(1 + 2/100) = 0.009$ . For 50% growth,  $\text{LOG}(1 + 50/100) = 0.176$ . Even for 180% growth,  $\text{LOG}(1 + 180/100) = 0.447$ , which is a fraction. Related log series of the exponential growth series then represent constant additions (accumulation) of  $\log(f)$  [i.e.  $\text{LOG}_{10}(f)$ ] from an initial base of  $\log(B)$  [i.e.  $\text{LOG}_{10}(B)$ ]. While this related log series grows ever larger, mantissa on the other hand is constantly oscillating between 0 and 1 as it takes many small steps forward, then suddenly one large leap backwards whenever log overflows an integer, and so forth. This is so since mantissa is obtained by constantly removing the whole part of the log (*whenever series  $\geq 1$ , that is, whenever  $\log(\text{series})$  is positive or zero, an assumption which can be taken for granted here*). Therefore, mantissa is clearly seen as being uniform on  $(0, 1)$  as more and more points of newly minted mantissa keep falling there on a variety of points, covering an ever increasing 'portion' of the entire  $(0, 1)$  range. Even though the process is truly deterministic, if one were to visually follow these 'rapid' mantissa additions onto  $(0, 1)$  space without getting severely dizzy, it would all seem quite random, disorganized, and highly chaotic, and it is precisely this nature of mantissa creation that yields uniformity to its final overall distribution!

As an example, we examine (typical) exponential series from base 3, with 30% growth rate, having factor  $f = 1.30$ , and with the implied  $\log(f) = 0.1139433$ . Figure 6.1 depicts part of that series in details. What should be carefully noted here is that mantissa always re-enters into  $(0, 1)$  interval at different or newer locations, namely at 0.047, 0.072, 0.098, and so on. This is a necessary condition for logarithmic behavior, as it guarantees that mantissa is well spread over the entire  $(0, 1)$  range in an even and uniform manner, covering all corners and segments. In a sense it guarantees that mantissa creation is random and unorganized in relationship to  $(0, 1)$  space, and thus that (almost) always **new mantissa** is being created.

Interestingly, such an indirect way of looking at exponential growth series (via their related log series) leads to the detection of some peculiar digital singularities. The argument above falls apart whenever the fraction  $\log(f)$  happens to be such that exactly  $M$  whole multiples of it add up to unity, as in the fractions 0.50, 0.25, 0.10, 0.125, 0.05, and so forth, in which case constant  $\log(f)$  additions lead to

<b>Series</b>	<b>Log</b>	<b>Mantissa</b>
3.0	0.477	0.477
3.9	0.591	0.591
5.1	0.705	0.705
6.6	0.819	0.819
8.6	0.933	0.933
11.1	1.047	0.047
14.5	1.161	0.161
18.8	1.275	0.275
24.5	1.389	0.389
31.8	1.503	0.503
41.4	1.617	0.617
53.8	1.730	0.730
69.9	1.844	0.844
90.9	1.958	0.958
118.1	2.072	0.072
153.6	2.186	0.186
199.6	2.300	0.300
259.5	2.414	0.414
337.4	2.528	0.528
438.6	2.642	0.642
570.1	2.756	0.756
741.2	2.870	0.870
963.6	2.984	0.984
1252.6	3.098	0.098
1628.4	3.212	0.212
2116.9	3.326	0.326

**Figure 6.1** Normal Logarithmic Exponential Series

re-entering  $(0, 1)$  always at the same point, and taking the same type of steps over and over again, focusing only on a few selected fortunate points having some very strong concentration (density), all the while ignoring all other points or sections on the interval  $(0, 1)$ . All this results in some quite uneven distribution on  $(0, 1)$  and thus yielding non-logarithmic digital distribution for the exponential series itself. Symbolically, the series is non-logarithmic and rebellious whenever there exists an integer  $M$  such that  $\log(f) \cdot M = 1$ . One should always be reminded that only uniformity of mantissa yields logarithmic behavior.

As an example, we examine one such anomalous series starting from base 8, growing at 77.8279% per period, and thus having factor  $f = 1.778279$  and the

<b>series</b>	<b>log</b>	<b>mantissa</b>
<b>8.0</b>	<b>0.903</b>	<b>0.903</b>
<b>14.2</b>	<b>1.153</b>	<b>0.153</b>
<b>25.3</b>	<b>1.403</b>	<b>0.403</b>
<b>45.0</b>	<b>1.653</b>	<b>0.653</b>
<b>80.0</b>	<b>1.903</b>	<b>0.903</b>
<b>142.3</b>	<b>2.153</b>	<b>0.153</b>
<b>253.0</b>	<b>2.403</b>	<b>0.403</b>
<b>449.9</b>	<b>2.653</b>	<b>0.653</b>
<b>800.0</b>	<b>2.903</b>	<b>0.903</b>
<b>1422.6</b>	<b>3.153</b>	<b>0.153</b>
<b>2529.8</b>	<b>3.403</b>	<b>0.403</b>
<b>4498.7</b>	<b>3.653</b>	<b>0.653</b>
<b>8000.0</b>	<b>3.903</b>	<b>0.903</b>
<b>14226.2</b>	<b>4.153</b>	<b>0.153</b>
<b>25298.1</b>	<b>4.403</b>	<b>0.403</b>
<b>44987.2</b>	<b>4.653</b>	<b>0.653</b>
<b>79999.7</b>	<b>4.903</b>	<b>0.903</b>
<b>142261.8</b>	<b>5.153</b>	<b>0.153</b>
<b>252981.2</b>	<b>5.403</b>	<b>0.403</b>
<b>449871.1</b>	<b>5.653</b>	<b>0.653</b>
<b>799996.3</b>	<b>5.903</b>	<b>0.903</b>
<b>1422616.6</b>	<b>6.153</b>	<b>0.153</b>
<b>2529809.3</b>	<b>6.403</b>	<b>0.403</b>
<b>4498706.7</b>	<b>6.653</b>	<b>0.653</b>

**Figure 6.2** Anomalous Non-Logarithmic Exponential Series

implied problematic  $\log(f) = \log(1.778279) = 0.25$  where exactly 4 (called  $M$ ) multiples of it add up to 1. Figure 6.2 depicts part of that series in detail. Mantissa always re-enters into  $(0, 1)$  interval at the same location, namely at 0.153, and then always takes the same subsequent 'long' steps, intentionally skipping 'numerous' points in between. Obviously such a state of affairs cannot result in any logarithmic behavior.

Yet, even for those rebellious rates, some comfort and relief can be found whenever  $\log(f)$  [the width of the steps by which log of the series advances along the log-axis] is sufficiently small compared with unit length, because then no matter how repetitive, peculiar, and picky mantissa chooses the points upon which to stamp on  $(0, 1)$ , each step is still so tiny that it has no choice but to walk almost all over the interval covering most corners and locations of  $(0, 1)$ . Simply put: the

creature is such that its legs are too short to be truly picky, so it cannot jump and skip much and is reduced to walking almost all over the interval, willingly or unwillingly. Only a walking creature with long legs can be effectively picky and successfully avoid certain segments lying on the ground. A giraffe can successfully avoid a 50-centimeter hole or gap on the ground, but a tiny ladybug cannot, no matter how carefully it walks.

Therefore, as the value of  $\log(f)$  decreases and becomes very small, digital configuration of these anomalous series begins to resemble monotonically decreasing digital configuration favoring low digits. This is so whenever  $\log(f)$  value is approximately below 0.1428 (namely  $1/7$ ). For even lower values of  $\log(f)$  the series begins to look more logarithmic-like. When  $\log(f)$  is very small, say, 0.01 (a rational  $1/100$ , designated  $1/M$ ), its spread over  $(0, 1)$  is fairly good and quite even so that its digit distribution is very near the logarithmic, even though upon closer examination it still stamps cautiously and discretely on the  $\log/\text{mantissa}$ -axis in even but tiny steps of 0.01 width each.

A table showing many of these anomalous series up to  $\log(f) = 0.005$  shall be generated. The two relationships  $f = (1+P/100)$  and  $\log(f)*M = 1$  (with  $M$  being an integer) imply that the argument above about anomalous rates translates into  $\text{LOG}_{10}(1+P/100) = 1/M$ , where  $M$  is an integer. Solving for  $P$  (the percent growth) by taking 10 to the power of both sides of the equation yields:

$$\begin{aligned} (1 + P/100) &= 10^{1/M} \\ P/100 &= 10^{1/M} - 1 \\ P\% &= 100*(10^{1/M} - 1) \quad M = 1, 2, 3, \text{ etc.} \end{aligned}$$

The table in Fig. 6.3 uses the above relationship to generate anomalous growth rates by varying integer  $M$  from 1 to 20, as well as evaluating  $M$  at 25, 35, 40, 50, 100, and 200. For each anomalous rate in the table an actual computer simulation (or rather, calculation) of the relevant exponential series is run, using the first 1000 elements, all starting from the initial (arbitrary) base value of 3. Digital results from such computer runs are displayed in the table, along with their associated SSD.

Even slight deviations from those anomalous rates above (*moving back to normal rates*) result in a near-perfect logarithmic behavior. For example, let us consider the rebellious rate  $100*(10^{(1/12)} - 1)\%$  series for  $M = 12$ , namely the growth 21.152765862859%. For this troubled series, an addition or subtraction of about 0.02% growth rate (i.e. about 21.1727% and 21.1327%) is enough to better its digital behavior and to arrive very close to the logarithmic. Well, this may be obvious since neither 21.1727% nor 21.1327% are rebellious; there exists no integral

M	LOG(F)	% Growth	1	2	3	4	5	6	7	8	9	SSD
1	1	900.000%	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	9155.8
2	0.5	216.228%	0.0	0.0	50.0	0.0	0.0	0.0	0.0	0.0	50.0	4947.6
3	0.333	115.443%	33.3	0.0	33.3	0.0	0.0	33.3	0.0	0.0	0.0	1701.8
4	0.25	77.828%	25.0	22.7	2.3	0.0	25.0	0.0	0.0	0.0	25.0	1063.3
5	0.2	58.489%	40.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	926.9
6	0.167	46.780%	16.6	16.6	16.7	16.7	0.0	16.7	0.0	0.0	16.7	619.8
7	0.143	38.950%	28.6	14.2	14.3	14.3	14.3	0.0	0.0	14.3	0.0	262.9
8	0.125	33.352%	25.0	24.5	0.5	12.5	12.5	0.0	12.5	0.0	12.5	424.9
9	0.111	29.155%	33.3	11.1	22.3	0.0	11.1	11.1	0.0	11.1	0.0	362.6
10	0.1	25.893%	30.0	10.0	20.0	10.0	10.0	0.0	10.0	0.0	10.0	236.7
11	0.091	23.285%	36.4	14.1	13.1	9.1	9.1	9.1	0.0	9.1	0.0	130.3
12	0.083	21.153%	24.9	16.6	16.8	8.4	8.4	8.3	8.3	0.0	8.3	97.4
13	0.077	19.378%	30.8	22.8	7.9	7.7	7.7	7.7	7.7	7.7	0.0	84.8
14	0.071	17.877%	28.4	20.9	7.7	14.4	7.2	7.2	0.0	7.1	7.1	103.6
15	0.067	16.591%	33.2	19.4	7.2	13.4	6.7	6.7	6.7	6.7	0.0	80.3
16	0.063	15.478%	31.0	12.4	12.6	12.6	6.3	6.3	6.3	6.3	6.2	43.5
17	0.059	14.505%	35.3	11.6	17.7	5.9	11.8	5.9	5.9	5.9	0.0	141.9
18	0.056	13.646%	27.5	16.5	16.8	5.6	11.2	5.6	5.6	5.6	5.6	56.6
19	0.053	12.884%	31.4	15.6	15.9	10.6	5.3	5.3	10.6	5.3	0.0	71.0
20	0.05	12.202%	30.0	15.0	15.0	10.0	10.0	5.0	5.0	5.0	5.0	21.2
25	0.04	9.648%	28.0	16.0	16.0	8.0	8.0	8.0	4.0	4.0	8.0	40.1
35	0.029	6.800%	28.1	16.8	14.5	8.7	8.7	5.8	5.8	5.8	5.8	13.1
40	0.025	5.925%	30.0	17.5	12.5	10.0	10.0	5.0	7.5	5.0	2.5	14.5
50	0.02	4.713%	30.0	16.0	14.0	10.0	8.0	6.0	6.0	4.0	6.0	8.8
100	0.01	2.329%	30.0	17.9	12.1	10.0	8.0	6.0	6.0	5.0	5.0	1.1
200	0.005	1.158%	30.0	17.5	12.5	10.0	8.0	6.5	6.0	5.0	4.5	0.2
Ben	=====	=====	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6	0.0

Figure 6.3 Anomalous Non-Logarithmic Exponential Growth Series — Basic Type

M satisfying  $P^0 = 100 \cdot (10^{1/M} - 1)$  for these two rates, yet just being somewhat close to a rebellious rate spells trouble to a small degree. For rates that are extremely close to a given rebellious one (yet not exactly equal), logarithmic behavior will not be obvious unless thousands or millions of elements are considered in the series, depending on (I) how close it is to the anomalous rate, and (II) the magnitude of the rebelliousness of the anomalous rate itself. For example, for  $M = 100$ , that is, for 2.3292992% growth rate, digit configuration is already very nearly logarithmic, and very little improvement is obtained by raising or lowering the rate slightly in order to escape rebelliousness.

Another ‘more direct’ vista of the digital characteristics of rebellious rates [without resorting to log or mantissa] is the empirical observation that for all the anomalous  $P^0$  growth rates above in Fig. 6.3, and only for those rates, there exists an integer  $N$  such that  $(1 + P/100)^N = 10$ . This last equation actually can also be

derived from the earlier expression  $\text{LOG}_{10}(1+P/100) = 1/M$  substituting  $N$  for  $M$ . This equality implies that applying  $(1+P/100)$  factor  $N$  times repeats the same element in the series except that its decimal place is shifted once to the right, and thus that its significand hasn't been altered. This in turn implies that the series itself repeats a list of  $N$  significands over and over again between IPOT points, precluding conformity to the logarithmic whenever  $N$  is small. For example, for the anomalous rate 77.828% in Fig. 6.3, there exists an integer  $N$ , namely 4, such that  $(1 + 77.828/100)^4 = 10$ . The emergence of these particular cumulative factors whose values are IPOT is cyclical, with each particular rate having a particular period, namely  $N$ . Clearly, the larger the value of  $N$  the more diluted is the net effect on overall digit distribution of the series, since a large value of  $N$  implies that this digital repetition does not happen very often. All this is nicely consistent with what was established earlier, namely the lessening in the severity of the deviations from the logarithmic for those rebellious series with lower growth rates (whose values of  $N$  are much larger, implying much longer periods and diluted effects).

Little reflection is needed to realize that  $N$  above (the  $N$ th cumulative factor) is simply  $M$ , the number of whole multiples of the fraction  $\log(f)$  adding up exactly to unity. That is, if exactly  $M$  whole multiples of the fraction  $\log(f)$  is unity, then the  $M$ th ( $N$ th) cumulative factor is then exactly 10, since returning to the very same mantissa means that digital configuration hasn't changed, and therefore that cumulative factor must be 10, or else perhaps another higher power of ten.

Anomalous series of the form described above, where the fraction  $\log(f)$  happens to be such that exactly whole multiples of it add up to unity, are actually just one particular type. More generally, whenever whole multiples of  $\log(f)$  add up any integer, be it 2, 3, or any other integral number, we encounter the same dilemma of having uneven mantissa distribution on  $(0, 1)$ , and thus non-logarithmic digital distribution for the exponential series itself. To recap, **General types** of anomalous exponential growth rates are found when exactly **T** whole multiples of the fraction  $\log(f)$  add up exactly to any integer **L**, and not just to unity. For example, 5 whole multiples of 0.4 yield exactly the value of 2 units of distance spanning  $\log/\text{mantissa}$  interval, hence its related 151.1886% growth series [ $0.4 = \text{LOG}_{10}(1+151.1886/100) = 2/5$ ] has a non-logarithmic digit distribution, its fifth cumulative factor is 100, its tenth cumulative factor is 10,000, and so forth. The general rule for any fraction  $\log(f)$ ,  $T$  as an integral multiple of that interval, and  $L$  an integral number, is as follows: whenever **the fraction  $\log(f)$  equals the rational number  $L/T$** , non-logarithmic behavior for the series itself is found.

L	T	LOG(F)	% Growth	1	2	3	4	5	6	7	8	9	SSD
2	5	0.4000	151.189%	40.0	0.0	20.0	20.0	0.0	0.0	20.0	0.0	0.0	926.9
2	7	0.2857	93.070%	28.5	28.0	0.6	14.3	14.3	0.0	0.0	14.3	0.0	497.7
2	9	0.2222	66.810%	33.3	11.1	22.3	0.0	11.1	11.1	0.0	11.1	0.0	362.6
2	13	0.1538	42.510%	30.8	21.6	9.1	7.7	7.7	7.7	7.7	7.7	0.0	64.2
2	17	0.1176	31.113%	35.2	16.1	13.3	5.9	11.8	5.9	5.9	5.9	0.0	80.5
2	23	0.0870	22.168%	26.1	17.3	13.1	13.0	4.4	8.7	4.3	4.4	8.7	63.6
2	25	0.0800	20.226%	28.0	16.0	16.0	8.0	8.0	8.0	4.0	4.0	8.0	40.1
2	47	0.0426	10.294%	29.4	16.8	12.9	10.8	8.6	6.4	6.5	4.3	4.3	4.3
2	67	0.0299	7.115%	30.0	17.4	12.1	9.0	9.0	6.0	6.0	4.5	6.0	4.8
2	71	0.0282	6.701%	29.4	16.8	13.1	9.9	8.4	7.0	5.6	4.2	5.6	3.8
2	120	0.0167	3.912%	29.6	17.6	12.0	10.2	8.5	6.8	5.1	5.1	5.1	1.9
2	214	0.0093	2.175%	29.7	16.2	14.0	10.0	9.0	6.7	5.4	5.4	3.6	6.9
2	344	0.0058	1.348%	31.0	15.0	13.2	10.2	7.8	7.2	6.0	5.4	4.2	8.9
2	657	0.0030	0.703%	29.7	17.3	13.9	9.4	7.8	6.6	5.7	5.1	4.5	2.4

Figure 6.4 Anomalous Exponential Series — General Type (L = 2, Increasing T)

That is, whenever  $\text{LOG}_{10}(1+P/100) = L/T$  with both L & T being positive integers, the series is rebellious. Solving for P in the above expression we get:  $P\% = 100*(10^{L/T} - 1)$ . For this general type of anomalous rates, the direct explanation blaming multiple integral powers of ten for its rebellious digital behavior is found in  $(1+P/100)^T = 10^L$ . The larger the value of T the less severe is the observed digital deviation. For T values over approximately 50, the behavior is quite nearly logarithmic regardless of what value L takes. For a given value of T, digital deviations are almost of the same magnitude for all values of L. Cumulative factors equal to integral powers of tens emerge cyclically with T as the period, and are of the forms  $10^L, 10^{2L}, 10^{3L}, 10^{4L}$ , and so forth. The table in Fig. 6.4 lists digital results of some such anomalous rates for various T values, with L fixed at 2. These simulations are of the first 1000 elements, all starting from the initial base value of 3. Other anomalous rates of the general type, from base 3 and having 1000 sequences, are shown in Fig. 6.5 for a wider variety of L and T values.

In Fig. 6.6 we examine actual (cumulative) factors of series, as opposed to log and mantissa point of view seen in Figs. 6.1 and 6.2. If the focus is on actual cumulative factors, then indirectly the focus is also on the actual series itself, ‘especially’ if the value of the base B is 1 in which case the series corresponds exactly to the cumulative factors. But even if the value of the base is not 1, the product of the base times the series of the cumulative factors is nothing but the series itself! The table in Fig. 6.6 demonstrates the difference in the patterns of cumulative factors (and thus the difference in the series themselves) between normal

L	T	LOG(F)	% Growth	1	2	3	4	5	6	7	8	9	SSD
3	4	0.7500	462.341%	25.1	0.0	25.1	0.0	24.8	0.0	0.0	0.0	25.1	1397.2
3	5	0.6000	298.107%	40.0	0.0	20.1	20.1	0.0	0.0	19.9	0.0	0.0	925.6
3	7	0.4286	168.270%	28.5	14.3	14.3	14.2	14.3	0.0	0.0	14.3	0.0	262.1
4	9	0.4444	178.256%	33.3	21.4	12.0	0.0	11.1	11.1	0.0	11.1	0.0	238.9
3	11	0.2727	87.382%	36.3	9.1	18.2	9.1	9.1	9.1	0.0	9.1	0.0	221.3
4	11	0.3636	131.013%	36.4	9.1	18.2	9.1	9.1	9.1	0.0	9.1	0.0	222.0
5	12	0.4167	161.016%	24.9	16.6	16.8	8.3	8.4	8.3	8.4	0.0	8.4	98.5
4	13	0.3077	103.092%	30.7	20.7	10.1	7.7	7.7	7.7	7.7	7.7	0.0	51.9
4	17	0.2353	71.907%	35.2	11.8	17.7	5.9	11.7	5.9	5.9	5.9	0.0	137.7
3	25	0.1200	31.826%	28.0	16.0	16.0	8.0	8.0	8.0	4.0	4.0	8.0	40.1
4	25	0.1600	44.544%	28.0	19.8	12.2	8.0	8.0	8.0	4.0	4.0	8.0	30.1
17	25	0.6800	378.630%	28.0	15.9	16.1	7.9	7.9	8.2	4.0	4.0	7.9	41.5
23	25	0.9200	731.764%	28.1	16.1	15.8	7.8	8.1	8.1	3.9	4.2	8.1	39.8
3	67	0.0448	10.860%	29.8	16.2	13.5	9.0	9.0	6.0	6.0	4.5	6.0	7.7
6	67	0.0896	22.900%	29.6	17.8	12.1	9.0	9.0	6.0	6.0	4.5	6.0	5.0
20	67	0.2985	98.842%	29.8	17.9	12.0	9.0	9.0	6.0	5.9	4.5	5.9	4.7
50	67	0.7463	457.531%	29.8	16.5	13.3	8.7	9.2	5.8	5.8	4.6	6.3	8.7
20	123	0.1626	45.412%	30.0	17.8	12.2	9.8	8.2	6.5	5.6	5.0	4.9	0.4
20	133	0.1504	41.376%	30.0	17.1	13.0	9.8	8.2	6.1	6.0	5.3	4.5	1.1
3	344	0.0087	2.028%	30.0	16.2	12.7	10.2	8.1	6.9	6.0	5.4	4.5	2.5
277	600	0.4617	189.512%	30.1	17.2	12.9	9.6	7.8	6.7	6.0	5.5	4.0	0.8
20	1223	0.0164	3.837%	29.4	17.3	12.9	10.1	8.3	6.8	5.7	5.0	4.5	1.1
277	3000	0.0923	23.690%	30.1	17.4	12.7	9.8	8.1	6.4	6.0	4.9	4.6	0.3
500	11200	0.0446	10.826%	30.1	17.6	11.8	9.9	8.1	7.2	5.4	5.4	4.5	1.1
747	13577	0.0550	13.506%	30.2	17.6	12.5	9.7	7.8	6.8	5.8	5.1	4.5	0.0

Figure 6.5 Other Anomalous Exponential Series — General Type (Increasing T)

and anomalous rates, explaining their corresponding logarithmic and rebellious behaviors. The first two (basic type) anomalous rates 29.154% and 58.489% are derived from  $100*(10^{1/9} - 1)$  and  $100*(10^{1/5} - 1)$ , respectively, with perfect repetition of the digital structure after each IPOT points. Of note here is the regularity in the emergence of IPOT factors, and their steadily increasing size one additional decimal place at a time. The third 40.000% series is chosen to represent a normal (and very typical) series with the usual near-perfect logarithmic behavior, absent any IPOT cumulative factors, and no repetition whatsoever. The fourth (general type) anomalous rate 93.070% is derived from  $100*(10^{2/7} - 1)$ .

We can confidently expect not merely most, but rather almost all exponential series to behave logarithmically. This is so for two reasons. Firstly, for a rate to be rebellious, the fraction  $\log(1+P/100)$  must equal exactly some rational number  $L/T$ , or at least be very close to it, and this is very rare. In fact, statistically speaking, the probability of obtaining a rational number when one is picked at random from

Cumulative Factor	29.155 %	58.489 %	40.000 %	93.070 %
The 1st cumulative factor	1.29	1.58	1.40	1.93
The 2nd cumulative factor	1.67	2.51	1.96	3.73
The 3rd cumulative factor	2.15	3.98	2.74	7.20
The 4th cumulative factor	2.78	6.31	3.84	13.89
The 5th cumulative factor	3.59	<b>10.00</b>	5.38	26.83
The 6th cumulative factor	4.64	15.85	7.53	51.79
The 7th cumulative factor	5.99	25.12	10.54	<b>100.00</b>
The 8th cumulative factor	7.74	39.81	14.76	193.07
The 9th cumulative factor	<b>10.00</b>	63.10	20.66	372.76
The 10th cumulative factor	12.92	<b>100.00</b>	28.93	719.69
The 11th cumulative factor	16.68	158.49	40.50	1389.50
The 12th cumulative factor	21.54	251.19	56.69	2682.70
The 13th cumulative factor	27.83	398.11	79.37	5179.47
The 14th cumulative factor	35.94	630.96	111.12	<b>10000.00</b>
The 15th cumulative factor	46.42	<b>1000.00</b>	155.57	19306.98
The 16th cumulative factor	59.95	1584.89	217.80	37275.94
The 17th cumulative factor	77.43	2511.89	304.91	71968.57
The 18th cumulative factor	<b>100.00</b>	3981.07	426.88	138949.55
The 19th cumulative factor	129.15	6309.57	597.63	268269.58
The 20th cumulative factor	166.81	<b>10000.00</b>	836.68	517947.47
The 21st cumulative factor	215.44	15848.93	1171.36	<b>1000000.00</b>
The 22nd cumulative factor	278.26	25118.86	1639.90	1930697.73
The 23rd cumulative factor	359.38	39810.72	2295.86	3727593.72
The 24th cumulative factor	464.16	63095.73	3214.20	7196856.73
The 25th cumulative factor	599.48	<b>100000.00</b>	4499.88	13894954.94
The 26th cumulative factor	774.26	158489.32	6299.83	26826957.95
The 27th cumulative factor	<b>1000.00</b>	251188.64	8819.76	51794746.79
The 28th cumulative factor	1291.55	398107.17	12347.67	<b>10000000.00</b>
The 29th cumulative factor	1668.10	630957.34	17286.74	193069772.89

Figure 6.6 Cumulative Factors (Series/Base) for Four Exponential Growth Series

any continuous set of real numbers is zero. And even though often financial rates, economics-related rates, and others are quoted rationally as fractions, this does not usually lead into any anomalous trap, because log is involved and it is quite rare that  $\log(1 + \text{Rational-Rate}/100) = (\text{Rational } L/T)$  could still hold. Nevertheless, merely being very close to a rebellious rate is problematic and logarithmic behavior is disrupted unless a truly large number of elements is considered to overcome the closeness to such a rate (depending on the intensity of the rebellionness of the anomalous rate and on the closeness to it.) Secondly, even if a given rate is rebellious or close to one, there is a cap on deviation given by the value of T. As argued and seen earlier, whenever T is roughly over 100, deviation from the logarithmic is fairly small.

As an empirical check on the theoretical reasoning given in this chapter, a computer program was run to check on digital behavior of exponential growth series.

It simulated series from 1% growth all the way to 600% growth in increments of 0.01%, checking the series for their digital behavior. Strong confirmation is found and the empirical result is just as theoretically expected, even though it may sound puzzling or paradoxical to the social scientist: Any **deviant** digital behavior differing from the logarithmic is always exclusively associated and correlated with the **rationality** of its related  $\log(f)$  fraction. There seems to be nothing else adversely affecting digital behavior, except for the argument given in this chapter.

In Chapter 20 on multiplication processes the reader was reminded that exponential growth series are not free of units or scale, but rather depend on the length of the time used in defining the period, and that any high growth that is quoted using some fixed period of time (say, a year) could in principle be quoted as lower growth if defined over a shorter period of time (say, a month). Let us see how all this relates to anomalous rates without any contradictions. **An annual rebellious rate implies also some rebelliousness on the part of its equivalent monthly rate as well, albeit with much reduced intensity.** Equivalency implies that  $(1+YR/100) = (1+MT/100)^{12}$  hence  $\text{LOG}(1+YR/100) = \text{LOG}[(1+MT/100)^{12}]$ , so that  $\text{LOG}(1+YR/100) = 12 * \text{LOG}(1+MT/100)$ . Therefore if  $L/T = \text{LOG}(1+YR/100)$  where  $L/T$  is rational, then  $L/T = 12 * \text{LOG}(1+MT/100)$ , so that  $L/(12 * T) = \text{LOG}(1+MT/100)$ . This implies that  $\log(f)$  of related monthly rate is also rational (12 is an integer!) and therefore rebellious as well. But deviation from the logarithmic for the monthly rate is less severe as it comes with a higher value of  $T$  (namely  $12 * T$ ).

In general, any basic type series satisfying  $(1+P/100)^N = 10$  points to another basic type of a lower rate, simply by writing it as  $(1+P/100)^{(1/R)*R*N} = 10$  or  $((1+P/100)^{1/R})^{*R*N} = 10$ , so that the  $R$ th root of  $(1+P/100)$  points to another basic type anomalous rate. Also, any basic type series satisfying  $(1+P/100)^N = 10$  points to another higher rate (general or basic) type series. We simply raise both sides of the above equation to the  $L$ th (integral) power yielding  $(1+P/100)^{NL} = 10^L$  or  $((1+P/100)^L)^N = 10^L$  so that the  $L$ th power of  $(1+P/100)$ , namely  $(1+P/100)^L$  points to another anomalous series, provided  $L$  is an integer. When  $L$  is larger than  $N$  it points to a general type. When  $L$  is smaller than  $N$ , the type depends on whether or not  $L$  is a proper factor of  $N$  (basic) or not (general). If  $L$  is a proper factor of  $N$ , say  $QL=N$ , and  $Q$  is an integer,  $1 < Q < N$ , then  $((1+P/100)^L)^N = 10^L$  can be written as  $((1+P/100)^L)^{QL} = 10^L$ , and upon taking the  $L$ th root of both sides we obtain  $((1+P/100)^L)^Q = 10$ , pointing to a basic anomalous series. A few examples of the above discussion are given by:  $1.066050 = \sqrt[9]{(1.778279)}$ , and

$1.136464 = \sqrt[3]{(1.291550)}$ . Also, the basic type rate of 1.110336 points to  $1.110336^2 = 1.23284$  (basic type, 2 is a factor of 22) and to  $1.110336^6 = 1.87381$  (general type, 6 is not a factor of 22).

The mathematician Ralph Raimi who has eloquently written some of the first mathematically rigorous articles on Benford's Law in the 60s and 70s has briefly mentioned the existence of such anomalous series, using the term 'reentrant series'. The author is still quite content to re-invent this old wheel and to add a few new features to it.

In contrast with the digital pitfalls of deterministic exponential growth series, multiplicative random walk processes on the other hand are always nicely logarithmic. Such processes can be thought of as having a non-constant (random) growth rate [namely factor  $f$  being a random variable]. Here, we never stumble upon the perils of fractions whose multiples always add exactly to an integral value on the log-axis in a consistent manner. Repeated additions of random fractional-log values result in covering the entire  $(0, 1)$  mantissa space evenly and 'fully' whenever there are plenty of such fractional accumulations.

## SUPER EXPONENTIAL GROWTH SERIES

---

Classic exponential growth series such as  $\{B, Bf^1, Bf^2, Bf^3, Bf^4, \dots, Bf^N\}$  are characterized by having a constant  $f$  multiplicative factor  $\{f, f, f, \dots, f\}$ , with  $f > 1$ .

Super exponential growth series is a case where the factors themselves are growing exponentially and where logarithmic behavior is found in spite of the rather odd nature of such series! The series of the growing factors themselves is defined as

$\{f^1, f^2, f^3, f^4, f^5, \dots, f^N\}$ , where  $f$  stands for the initial factor as well as for the factor by which the factors themselves are growing. The series itself is then written in terms of its construction term by term as:

$$\{B, B(f^1), Bf^1(f^2), Bf^1f^2(f^3), Bf^1f^2f^3(f^4), Bf^1f^2f^3f^4(f^5), \dots, Bf^1f^2f^3f^4f^5 \dots (f^N)\}$$

$$\{B, Bf^1, Bf^3, Bf^6, Bf^{10}, Bf^{15}, \dots, Bf^{(N*N + N)/2}\}$$

This is essentially of the same format of the classic exponential growth series, but instead of having the exponents of the factors increasing sequentially simply as in 1, 2, 3, 4, etc. they are expanding more rapidly as in 1, 3, 6, 10, 15, etc. (or equivalently stated: that the differences in exponents are increasing sequentially as in 1, 2, 3, 4, and so forth, hence this super series shall be referred to as 'monotonic differences'). In one particular computer simulation example, base  $B$  is arbitrarily chosen as 100, and the factor of the factors  $f$  is chosen as 1.0008. The first-digits distribution of this super growth series considering its first 1328 elements is  $\{31.4\%, 16.8\%, 12.3\%, 9.9\%, 7.4\%, 5.6\%, 5.9\%, 5.6\%, 5.0\%\}$ . Its SSD is a rather low value of 4.3, signifying a great deal of closeness to the logarithmic. The difficult issue in creating and dealing with such super exponential growth series with the aid of the computer is the quick explosion upwards toward some very large numbers which the machine can't deal with. This upward explosion causes digital distribution to appear as if it is deviating from the logarithmic, since an insufficient number of elements are considered by the computer, masking true (logarithmic)

results. When the factor of the factors  $f$  is larger than, say, 1.01, rapid upward explosion limits the number of calculated elements to no more than about 400 on a normal personal computer. On the other hand, when the factor of the factors  $f$  is very small and near 1 (say,  $f < 1.0001$ ) digits are nearly frozen for too long at the same digital configuration of the base  $B$ , since no significant changes in the value of the elements occur initially as the super series 'progresses' on the computer screen. A little reflection is needed to realize that since  $f > 1$ , **the super growth series speeds up along the log-axis**, (i.e. distances between consecutive log values of the super series are increasing) and therefore no argument can be made here about evenly spaced walk along the log-axis, progressing by a constant value of  $\log(f)$  at a time, and leading eventually to uniformity of mantissa as was seen in the previous chapter regarding standard exponential growth series. Yet the super series appears to be nearly logarithmic! The fascination generated by super growth series stems from the empirical evidence that these series are exactly or approximately logarithmic regardless of the particular choices of the base or the factor of the factors, as long as plenty of elements can be considered on a powerful computer which doesn't suffer much from that upward explosion and could calculate some very long such series for tens of thousands of elements. A better way of going about it on the computer is following the series by way of related log series development, dealing with much lower values as opposed to following the actual values of the series themselves which are of very high values (in other words, simulating related log values directly). The rationale for the logarithmic behavior of super exponential growth series is that it speeds up along the log-axis in a random and haphazard fashion as far as mantissa is concerned. In other words, it's logarithmic because the series trams upon the log-axis in a disorganized and uneven manner between log integers, resulting in the accumulation of all sorts of mantissa values. This in turn guarantees that mantissa is approximately uniform since the process gives no preference to any type or sub-set of mantissa space, but rather keeps picking truly 'random' mantissa values. On the other hand, any organized and careful march along the log-axis coordinated between log integers leads to non-logarithmic behavior, as was seen in the case of anomalous exponential growth rates, where such structure in the walk along the log-axis results in rebellious non-logarithmic leading-digits behavior.

Surely many other rapidly growing  $f$  exponents series with similar formats may also be constructed, such as in utilizing the Fibonacci series 1, 1, 2, 3, 5, 8, etc. as those exponents of the  $f$  series for example, corresponding to another super series  $\{B, Bf^1, Bf^1, Bf^2, Bf^3, Bf^5, Bf^8, \dots\}$ . Without computer simulations as

confirmation and without any mathematical proofs, the conceptual consideration mentioned above regarding mantissa accumulation clearly points to full compliance with Benford’s Law here as well. Another even more extreme case may be considered when  $f$  exponents themselves are exploding forwards exponentially, constantly doubling, such as in say 1, 2, 4, 8, 16, 32, and so forth, corresponding to the series  $\{B, Bf^1, Bf^2, Bf^4, Bf^8, Bf^{16}, Bf^{32}, \dots\}$  which grows extremely rapidly! Needless to say such super explosion in values renders the series quite problematic on the typical computer (unless  $f$  is carefully selected to be just over 1 but not too close to it). In any case, logarithmic behavior here should not be in doubt. Both series, the Fibonacci-like and the double-exponent one, march (and speed up) haphazardly along the log-axis with no organization or order whatsoever in relation to integral log points (i.e. mantissa), guaranteeing logarithmic behavior for the series themselves.

Surely there are infinitely many ways to construct other super exponential series. Let us summarize the four series mentioned in this chapter, along with the limit of the ratio of consecutive exponents of their factors, namely the limit of the ratio  $\text{LOG}(f_{N+1})/\text{LOG}(f_N)$  as  $N$  approaches infinity:

- Classic series:  $\{B, Bf^1, Bf^2, Bf^3, Bf^4, Bf^5, \dots\}$   $L(f_{N+1})/L(f_N) \rightarrow 1$
- Monotonic diff.:  $\{B, Bf^1, Bf^3, Bf^6, Bf^{10}, Bf^{15}, \dots\}$   $L(f_{N+1})/L(f_N) \rightarrow 1$
- Fibonacci-like:  $\{B, Bf^1, Bf^1, Bf^2, Bf^3, Bf^5, Bf^8, \dots\}$   $L(f_{N+1})/L(f_N) \rightarrow 1.618$
- Double expon.:  $\{B, Bf^1, Bf^2, Bf^4, Bf^8, Bf^{16}, Bf^{32}, \dots\}$   $L(f_{N+1})/L(f_N) \rightarrow 2$

In spite of false appearances, both classic exponential as well as monotonic differences super series come with an additive-like exponent structure while the Fibonacci-like as well as double-exponent series come with a multiplicative-like exponent structure. [The Fibonacci quickly converges to an exponential-like golden ratio 1.618 growth series]. Hence for this short list of four series, perhaps only the Fibonacci-like and double-exponent series are the ones truly deserving the name ‘super exponential growth series’.

The examples in this chapter further demonstrate the tenacity and prevalence of Benford’s Law, as it appears valid even in such atypical cases and situations. On a more profound level though, all four series considered here, classic exponential series as well as super exponential series, not only represent the same concept, but all operate indeed under a singular, common, and identical mechanism. That mechanism is simply tramping upon the log-axis in an unorganized and haphazard manner not coordinated or repeated in relation to integral log values, culminating in uniformity of mantissa and the resultant logarithmic behavior.

## HIGHER-ORDER LEADING DIGITS

---



---

Benford's Law does not explicitly state the unconditional probabilities for second-, third-, and higher-order leading digits. Instead this is indirectly inferred by exhausting all possibilities of digit combinations. Surely it is possible to infer all aspects of Benford's Law from its general statement  $\text{Probability}(\text{significand} \leq S_0) = \text{LOG}_{10}(S_0)$ , but algebraically it is a bit complex. The law states directly how first digits are distributed, namely  $\text{LOG}(1 + 1/d)$ . The law also states directly how the first-two-digits combinations are distributed, namely  $\text{LOG}(1 + 1/pq)$ . By putting these two explicit algebraic expressions together one may obtain the probability of digit  $k$  leading the second order, regardless of what digit is leading first order, namely the unconditional probability of  $k$  in the second order. For digit  $k$  to lead second order, it could occur whenever 1, or 2, or 3, ... , or 9 are leading first, hence the need to add the nine distinct **first-two-digits** probabilities of  $1k, 2k, 3k, \dots, 9k$ .

### Unconditional second-order probability:

$P(\text{2nd digit} = k \mid \text{any 1st digit}) = \sum \log(1 + 1/Dk)$  summed over all  $D \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

For example, the chance that digit 8 is unconditionally leading second order is given by the sum:

$$\begin{aligned} & \log(1 + 1/18) + \log(1 + 1/28) + \log(1 + 1/38) + \log(1 + 1/48) + \\ & \log(1 + 1/58) + \log(1 + 1/68) + \log(1 + 1/78) + \log(1 + 1/88) + \\ & \log(1 + 1/98) = \mathbf{0.08757} \end{aligned}$$

Similarly, the probability that digit  $m$  leads third order, regardless of what digits are in the first and second places, is calculated by adding the probabilities of all possible relevant combinations. For digit  $m$  to lead third order, it occurs whenever **first-three-digits** combinations are:  $10m, 11m, 12m, 13m, \dots, 96m, 97m, 98m, 99m$ .

**Unconditional third-order probability:**

$P(\text{3rd digit} = m \mid \text{any 1st, any 2nd}) = \sum \sum \log(1 + 1/DK_m)$  summed over all  $D \in \{1,2,3,4,5,6,7,8,9\}$  and  $K \in \{0,1,2,3,4,5,6,7,8,9\}$

For example, the chance that digit 8 is unconditionally leading third order is given by the sum of the following 90 terms:

$$\begin{aligned} & \log(1 + 1/108) + \log(1 + 1/118) + \log(1 + 1/128) + \dots + \log(1 + 1/198) + \\ & \log(1 + 1/208) + \log(1 + 1/218) + \log(1 + 1/228) + \dots + \log(1 + 1/298) + \\ & + \dots + \\ & \log(1 + 1/808) + \log(1 + 1/818) + \log(1 + 1/828) + \dots + \log(1 + 1/898) + \\ & \log(1 + 1/908) + \log(1 + 1/918) + \log(1 + 1/928) + \dots + \log(1 + 1/998) \\ & = \mathbf{0.09864} \end{aligned}$$

Fourth-order unconditional distributions are similarly calculated by summing all relevant **first-four-digits** expressions of  $\log(1 + 1/pqrs)$ , although there is very little interest in pursuing probabilities of such high-order digit distribution as it is nearly uniform and dispenses the proportion of nearly 10% for each digit.

The calculations above refer to the **unconditional** probabilities. For example, the value of 0.08757 above represents the unconditional probability that digit 8 occurs on the second place regardless of what digit is occupying the first place. On the other hand, given that the first digit is  $d$ , then the **conditional** probability that  $k$  is the second digit is given by:

**Conditional probability  $k$  is 2nd given  $d$  is 1st:**

$$\begin{aligned} P(\text{2nd digit} = k \mid \text{1st digit} = d) &= P(\text{first-two digits} = dk) / P(\text{1st digit} = d) \\ &= \log(1 + 1/dk) / \log(1 + 1/d) \end{aligned}$$

For example, given that first digit is 3, then the conditional probability that 8 is the second digit is calculated as  $P(\text{first-two digits} = 38) / P(\text{1st digit} = 3) = \log(1 + 1/38) / \log(1 + 1/3) = \mathbf{0.09029}$ .

Also, the conditional probability that 0 is the second digit given that first digit is 3 is  $P(\text{first-two digits} = 30) / P(\text{1st digit} = 3) = \log(1 + 1/30) / \log(1 + 1/3) = \mathbf{0.11398}$ .

The table in Fig. 6.7 gives the conditional probabilities of second-order digits for any given first-order occurrence [calculated as in the above two examples].

Given that 1st digit is	Probabilities for 2nd digit:									
	0	1	2	3	4	5	6	7	8	9
1	13.8	12.6	11.5	10.7	10.0	9.3	8.7	8.2	7.8	7.4
2	12.0	11.5	11.0	10.5	10.1	9.7	9.3	9.0	8.7	8.4
3	11.4	11.0	10.7	10.4	10.1	9.8	9.5	9.3	9.0	8.8
4	11.1	10.8	10.5	10.3	10.1	9.8	9.6	9.4	9.2	9.1
5	10.9	10.7	10.4	10.3	10.1	9.9	9.7	9.5	9.4	9.2
6	10.7	10.5	10.4	10.2	10.1	9.9	9.8	9.6	9.5	9.3
7	10.6	10.5	10.3	10.2	10.1	9.9	9.8	9.7	9.5	9.4
8	10.5	10.4	10.3	10.2	10.0	9.9	9.8	9.7	9.6	9.5
9	10.5	10.4	10.3	10.2	10.0	9.9	9.8	9.7	9.6	9.5

Figure 6.7 Conditional Probabilities of Second Order Given First Order

Therefore second-order digits are more skewed (in favor of low digits) whenever the first digit is low, and are more equal whenever the first digit is high. This is only one example of the dependencies between the orders of the digits. Knowing that the first digit is low implies that chances are relatively higher a bit that the second digit is low as well. Knowing that the first digit is high implies that chances are relatively a bit higher that the second digit is high as well.

Let us reverse the order of dependency. Given that  $k$  is unconditionally the second digit, then the **conditional** probability that  $d$  is the first digit is given by:

**Conditional probability  $d$  is 1st given  $k$  is 2nd:**

$$\begin{aligned}
 &P(1st\ digit = d \mid 2nd\ digit\ is\ unconditionally\ k) = \\
 &P(first\ two\ digits = dk) / P(2nd\ digit\ is\ unconditionally\ k) = \\
 &\log(1 + 1/dk) / [\log(1 + 1/1k) + \log(1 + 1/2k) + \log(1 + 1/3k) + \dots + \log(1 + 1/9k)]
 \end{aligned}$$

For example, given that the second digit is unconditionally 4, then the condi-  
tional probability that the first digit is 7 is given by:

$$\begin{aligned}
 &\log(1 + 1/74) / [\log(1 + 1/14) + \log(1 + 1/24) + \log(1 + 1/34) + \dots + \log(1 + 1/94)] = \\
 &0.0058 / [0.0300 + 0.0177 + 0.0126 + 0.0098 + 0.0080 + 0.0067 + 0.0058 + 0.0051 + \\
 &0.0046] = \\
 &= \mathbf{0.0581}.
 \end{aligned}$$

Given that 2nd digit is	Probabilities for 1st digit:								
	1	2	3	4	5	6	7	8	9
0	34.6	17.7	11.9	9.0	7.2	6.0	5.1	4.5	4.0
1	33.2	17.7	12.1	9.2	7.4	6.2	5.3	4.7	4.2
2	31.9	17.7	12.3	9.4	7.6	6.4	5.5	4.8	4.3
3	30.8	17.7	12.4	9.6	7.8	6.6	5.7	5.0	4.5
4	29.9	17.7	12.6	9.7	7.9	6.7	5.8	5.1	4.6
5	29.0	17.6	12.7	9.9	8.1	6.9	6.0	5.3	4.7
6	28.2	17.6	12.7	10.0	8.2	7.0	6.1	5.4	4.8
7	27.5	17.5	12.8	10.1	8.4	7.1	6.2	5.5	4.9
8	26.8	17.4	12.9	10.2	8.5	7.2	6.3	5.6	5.0
9	26.2	17.3	12.9	10.3	8.6	7.4	6.4	5.7	5.1

**Figure 6.8** Conditional Probabilities of First Order Given Second Order

The table in Fig. 6.8 gives the conditional probabilities of first-order digits for any given second-order occurrence.

Therefore first-order digits are relatively more skewed than the unconditional logarithmic (in favor of low digits) whenever second digit is low, and are relatively a bit more equal whenever second digit is high. This is another example of the dependencies between the orders of the digits. Knowing that second digit is low implies that chances are relatively higher a bit that first digit is low as well. Knowing that second digit is high implies that chances are relatively higher that first digit is high as well. All this also applies to third and higher orders. In that sense, digit distributions of the various orders depend on one another. In other words:

**There is a positive correlation in rank (high/low) between the orders.**

In order to obtain again the unconditional probabilities of second-order digits directly from the conditional table of Fig. 6.7, one should not take the simple average by column, since first-order digits do not occur with equal probabilities, rather the Benford-weighted average column by column should be calculated. For

Digit	0	1	2	3	4	5	6	7	8	9
1st Order		30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
2nd Order	12.0	11.4	10.9	10.4	10.0	9.7	9.3	9.0	8.8	8.5
3rd Order	10.18	10.14	10.10	10.06	10.02	9.98	9.94	9.90	9.86	9.83
4th Order	10.02	10.01	10.01	10.01	10.00	10.00	9.99	9.99	9.99	9.98

Figure 6.9 Unconditional Probabilities of First, Second, Third, and Fourth Orders

example, the unconditional probability that 5 is the second-order digit is calculated as:

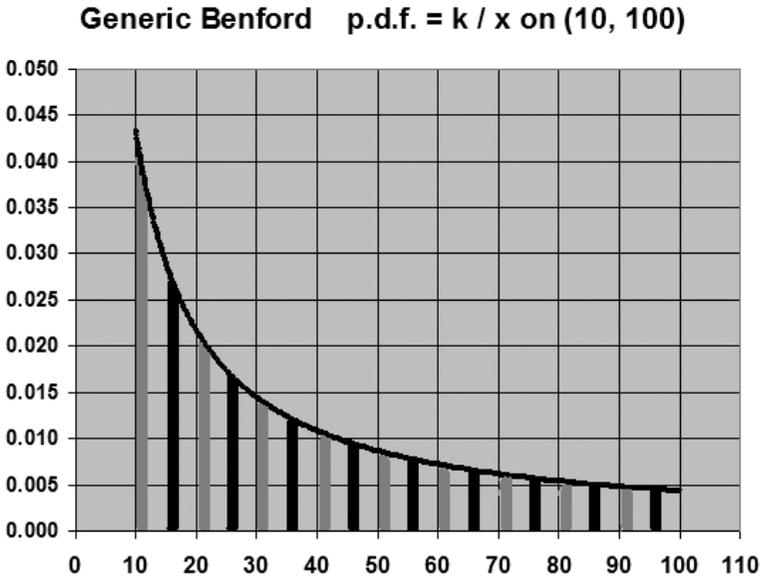
$$\begin{aligned}
 & P(\text{2nd digit} = 5 \mid \text{1st digit} = 1) * \log(1 + 1/1) + \\
 & P(\text{2nd digit} = 5 \mid \text{1st digit} = 2) * \log(1 + 1/2) + \\
 & P(\text{2nd digit} = 5 \mid \text{1st digit} = 3) * \log(1 + 1/3) + \\
 & + \dots + \\
 & P(\text{2nd digit} = 5 \mid \text{1st digit} = 9) * \log(1 + 1/9) \\
 & = 0.301 * 0.0931 + 0.176 * 0.0967 + 0.125 * 0.0979 + 0.097 * 0.0985 + \\
 & 0.079 * 0.0988 + 0.067 * 0.0990 + 0.058 * 0.0992 + 0.051 * 0.0993 + \\
 & 0.046 * 0.0994 = \mathbf{0.0966}
 \end{aligned}$$

The table in Fig. 6.9 provides the unconditional probabilities of first-, second-, third-, and fourth-order distributions.

Does Benford's Law imply that digit 1 occurs in computer files for example much more often than other digits, so much so that in the future, efficient engineering design should take advantage of this odd distribution? What hampers Benford's Law in this context is the fact that only the first digits are so highly skewed, while the higher orders (which are more numerous) are close to equality, hence overall effect on digit proportions is quite mild. Under the assumptions that all the numbers in a given data set are made of exactly four digits, then overall frequency in the usage of digits, regardless of order and position, can be easily calculated as the simple average of all the four order distributions shown in Fig. 6.9 digit by digit, yielding the relatively mild digital proportion of {8.0%, 15.4%, 12.2%, 10.7%, 9.9%, 9.4%, 9.0%, 8.7%, 8.4%, 8.2%}. Under the assumptions that all the numbers in a given data set are made of exactly seven digits, then overall frequency in the usage of digits regardless of order or position is calculated by the simple averaging of all the seven orders, yielding {8.9%, 13.1%, 11.2%, 10.4%, 10.0%, 9.7%, 9.4%, 9.2%, 9.1%, 9.0%}. Hence 1

is the most frequent digit, and 0 is the least frequent digit, while the effect of Benford's Law on overall keyboard usage or digital storage in computers for example is not as strongly skewed in favor of the low digits as is the case in the occurrences of the first digits where digit 1 is almost seven times more likely to occur than digit 9. This is so since higher orders (which predominate much more than the first order in seven-digit-long numbers) equalize and moderate overall distribution due to their own mild and less skewed configurations near digital equality. In any case, the fact that digits have unequal occurrences could be of interests to computer manufacturers such as IBM, INTEL, and others for the efficient engineering in the design of their products. Quoting what Theodore Hill writes on the issue: "It is possible to build computers whose designs capitalize on knowing the distribution of numbers they will be manipulating. If 9s are much less frequent than 1s then it should be possible to construct computers which use that information to minimize storage space, or to maximize rate of output printed. A good analogy is a cash register drawer. If the frequency of transactions involving the various denominations of bills is known, then the drawer may be specially designed to take advantage of that fact by using bins of different sizes. In fact, German mathematician Peter Schatte has determined that based on an assumption of Benford input, the best computer design that minimizes expected storage space (among all computers with binary-power base) is base 8, and other researchers are currently exploring use of logarithmic computers to speed calculations."

The disparity between the orders of digits, the decrease in magnitude of skewness as we consider higher and higher orders, as well as the dependencies between the orders can all be demonstrated and reasoned using the insight gained in earlier sections of the book and convincingly illustrated in the chart of Fig. 6.10 which depicts  $k/x$  density curve defined over an interval between the two adjacent IPOT numbers 10 and 100. The chart represents one very important type of the generic logarithmic distribution, a Benford prototype of sorts (even though it is restricted to a particular interval), therefore conclusions drawn from it may be considered quite general. True, real-life random data typically span multiple IPOT values, and it never falls off steadily as in the  $k/x$  case (since development pattern dominates), yet in the aggregate, over many IPOT sub-intervals, it does fall off as in  $k/x$ , hence such an analysis of the  $k/x$  case is quite relevant. In addition, any claim that presenting this particular curve over (10, 100) might be arbitrary, and that other curves could also serve as the density here, is easily refuted by Proposition III which lends it uniqueness, stating that only  $k/x$  curve is logarithmic over such an interval.



**Figure 6.10** Visualization of Higher Orders Existence, Skewness, and Dependency

We first note that the entire range of (10, 100) contains nine different **sub-regions**  $\{(10, 20), [20, 30), [30, 40), \dots, [90, 100)\}$ , each dominated by a different first digit.

The rectangular-like areas shaded light gray are those where second-order digits are exclusively of digit 0, such as (10, 11), [20, 21), [30, 31), and so forth. Those shaded dark black are where digit 5 leads second order, such as [15, 16), [25, 26), [35, 36), and so forth. Within each of the nine sub-regions, the gray areas of 0 are taller than their black sisters of 5, and thus contain more area, and more probability. This is why second order is also skewed in favor of low digits! Clearly all lower second-order digits (not just 0 vs. 5) have the advantage here over higher ones just as they have regarding first order, a fact demonstrated with taller pdf curve for lower second digits compared with shorter pdf curve for higher second digits, and all this is true for each of the nine sub-regions of the first order! This is so since the graph of  $k/x$  is continuously falling, and not only between regions regarding first order, but also within those regions and regarding second order. The consistent fall in the curve guarantees that lower digits are consistently favored over higher digits no matter what part of it (i.e. order) is considered.

Yet, within each of these nine first-order sub-regions, **local** curvature which pertains to the second order is not as concave or sharp compared with the entire

**global** (10, 100) interval pertaining to the first order which falls off quite dramatically. For third-order leading digits, the comparison is made on even shorter-size intervals, implying that curve is even flatter locally, and thus third-order digits experience more equality. For example, digit 0 leads third order on the tiny sub-sections of  $\{(10.0, 10.1), [11.0, 11.1), [12.0, 12.1), \dots, [98.0, 98.1), [99.0, 99.1)\}$ . Clearly the effects of the steady fall in density is diminishing for each higher order since that **global curvature appears relatively flatter locally** on shorter sub-sections, therefore digital skewness is diminishing for the higher orders, culminating in near digital equality for the fifth and higher orders where curve appears nearly flat locally.

A third insight can be obtained regarding the **dependencies between the orders**. For example, the fact that second order depends on first order can be clearly visualized by the reduction in the curvature of the curve within each of the nine sub-regions of first order as focus shifts to higher first-order digits. The curvature becomes steadily milder (curve flatter) as focus shifts from region (10, 20), to region [20, 30), to [30, 40), and so forth, culminating in a much flatter curve in the region [90, 100). Clearly, second-order inequality among the digits is milder for numbers with high first-order digits such as 9 [where curve is flatter], and it is more extreme for numbers with low first-order digits such as 1 [where curve is quite steep]. The table in Fig. 6.7 of conditional second order confirms this.

## DIGIT DISTRIBUTIONS ASSUMING OTHER BASES

---

An important generalization of Benford's Law relates to the ability to state it as a singular algebraic expression covering all bases for any positional number system in use. Had the law required different expressions for different bases, or if it could apply only for some bases such as 10 for earthlings say, but not for other bases, its reputation would suffer and doubts would abound. As it happens, Benford's Law is valid for any base, not only 10, and a simple adjusting factor is added to accomplish this important generalization. The probability for any base  $B$  of the first leading digit  $d$  (in the range 1 to  $B-1$ ) is  **$P[d \text{ is first}] = \text{LOG}_{10}(1 + 1/d)/\text{LOG}_{10}(B)$** . Other formulae such as those for higher-order significant digits or for the general law form are carried over to other bases with the same adjustment factor of  $1/\text{LOG}_{10}(B)$ . Applying the logarithmic identity  $\text{LOG}_A X = \text{LOG}_B X / \text{LOG}_B A$  to convert this expression directly into its base language, we get:

$$P[d \text{ is first}] = \text{LOG}_{10}(1 + 1/d)/\text{LOG}_{10}(B)$$

$$P[d \text{ is first}] = [\text{LOG}_B(1 + 1/d)/\text{LOG}_B(10)]/[\text{LOG}_B(B)/\text{LOG}_B(10)]$$

$$P[d \text{ is first}] = [\text{LOG}_B(1 + 1/d)/\text{LOG}_B(10)]/[1/\text{LOG}_B(10)]$$

$$\mathbf{P[d \text{ is first}] = \text{LOG}_B(1 + 1/d)}$$

The general law regarding **all orders**, can then incorporate **all bases** as well, and it is expressed as:

$$\mathbf{\text{Probability}(\text{significant} \leq S_0) = \text{LOG}_{\text{BASE}}(S_0)}$$

The elegance and conciseness in this result are striking, having the same form for all bases, and never even mentioning or involving that peculiar number **10** — significant only on that irrelevant planet called Earth on the very edge of the galaxy — thereby lending the expression universality. For binary numbers (base 2) the probability of digit 1 being the first leading digit is 100% since all leading digits of binary numbers are necessarily 1, in which case Benford's Law becomes a tautology. The table in Fig. 6.11 shows digital proportions for five different bases, as well as the ratios of the two extreme digits, expressing a measure of the dichotomy between the occurrences of the lowest and the highest digits. This shows that the higher the base the more severe is the dichotomy between the highest and lowest digit. The chart in Fig. 6.12 depicts digital configurations for four different base systems.

Digit	Base 4	Base 6	Base 10	Base 14	Base 21
1	50.0%	38.7%	30.1%	26.3%	22.8%
2	29.2%	22.6%	17.6%	15.4%	13.3%
3	20.8%	16.1%	12.5%	10.9%	9.4%
4		12.5%	9.7%	8.5%	7.3%
5		10.2%	7.9%	6.9%	6.0%
6			6.7%	5.8%	5.1%
7			5.8%	5.1%	4.4%
8			5.1%	4.5%	3.9%
9			4.6%	4.0%	3.5%
10				3.6%	3.1%
11				3.3%	2.9%
12				3.0%	2.6%
13				2.8%	2.4%
14					2.3%
15					2.1%
16					2.0%
17					1.9%
18					1.8%
19					1.7%
20					1.6%
<b>Ratio of Highest to Lowest:</b>	<b>2.4</b>	<b>3.8</b>	<b>6.6</b>	<b>9.4</b>	<b>14.2</b>

Figure 6.11 First-Digit Distribution and Digital Dichotomy for Various Bases

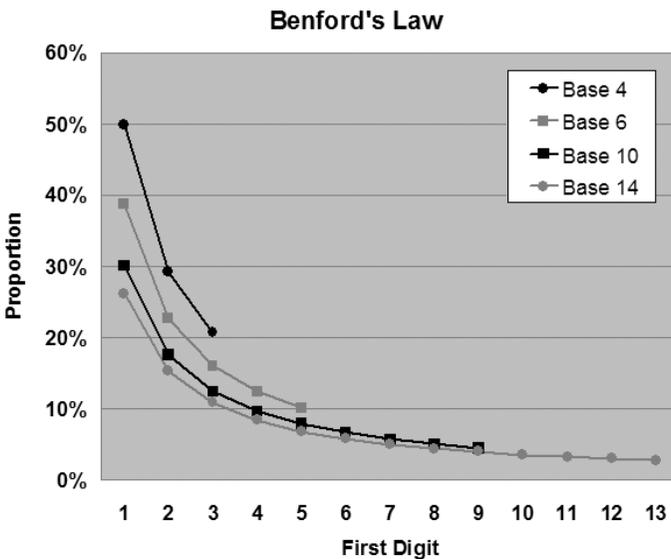


Figure 6.12 Chart of First-Digit Distribution for Various Bases



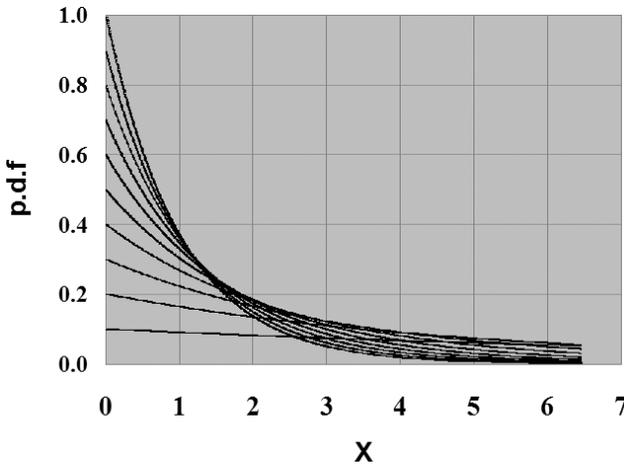
Uniform(0, U(0, U(0, U(0, U(0, U(0, U(0, U(0, U(0, 17)))))))))) is a nine-sequence chain of Uniforms, or **Flehinger's iterated scheme** in disguise! This is Flehinger's scheme except that Lower Bound is tied here to 0, while hers is fastened to 1. This is a crucial difference though, potentially yielding very different results for these similarly-looking schemes. It should be recalled that this short yet potent distance from 0 to 1 contains an infinite number of intervals between IPOT numbers, as well as an infinite number of related log negative integers! This chain of only nine (very finite) Uniforms converges rapidly, and simulations gave excellent first-digits results of {30.3, 17.8, 12.9, 9.5, 8.0, 6.5, 5.7, 4.9, 4.5}, with an extremely low SSD value of 0.4!

Uniform(0,  $1/(x \cdot \ln 10)$  defined over (1, 10)). This short but effective chain is nothing but **Frank Benford's own attempt at a proof** — in disguise! Simulations yielded the logarithmic almost exactly! Hence, Benford's curious and unsuccessful attempt at explanation can now be viewed as simply one very short yet powerful chain of distributions! More on Frank Benford's attempt at an explanation, and why it cannot serve as a valid proof, to be discussed in a later chapter in this section.

Limit  $N \rightarrow \infty$  Uniform(0, Uniform( $10^N$ ,  $10^{N+1}$ )),  $N$  is an integer. This 'chain limit' is nothing but **Stigler's law** in disguise! Simulation of the chain U(0, U(10000, 100000)) gave {24.3, 18.1, 14.4, 11.8, 9.5, 7.8, 6.3, 4.5, 3.3}.

The proper view here is the other way around, namely that Benford's attempt, Flehinger's scheme, and Stigler's law are nothing but chains of distributions in disguise. The generic idea of the chain can be thought of as the common thread going through all these schemes.

Conceptually it is easy to envision how we get approximately  $k/x$ -like or Lognormal-like shaped curve when distributions are chained, or at a minimum something with a definite tail to the right where curve is falling. Let us consider the two-sequence short chain **exponential(Uniform(0, 1))**, pointing to a set of exponentials, all starting at 0 (by default) and having progressively different focus on the  $(0, \infty)$  range as parameter varies on Uniform(0, 1). This chain does not converge to the logarithmic exactly, yet it is extremely close to the logarithmic, far more so than any single exponential with a fixed value for its parameter can aspire to be. It should be recalled that the exponential is defined over  $(0, +\infty)$ , and that the mean is  $1/\text{parameter}$ . We could attempt to envision overall density shape of the above chain by aggregating 10 typical realizations of exponential distributions serving as elements of the chain, and having parameters evenly spread over  $(0, 1)$ .



**Figure 6.13** Ten Fair Realizations from the Chain Exponential(Uniform(0, 1))

We choose the parametrical values of  $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$  as good representatives of this chain scheme. Figure 6.13 depicts those 10 individual exponential densities superimposed within a single chart.

Since all individual curves have common Lower Bound (namely zero, as for all exponentials) while Upper Bounds vary (depending on parameter), the result is a definite fall in the density of the aggregate, with a tail to the right. Although each individual exponential has such shape to begin with, namely a tail falling to the right, the chain has the more proper tail form and thus closer to the logarithmic. More dramatic results are obtained with **Uniform(0, Uniform(0, 1))** for example, which comes with a flat (tailless) curve for each individual uniform, yet gives rise to a curved falling tail in the density of the chain which aggregates many flat uniforms. This certainly reminds us of the crucial dichotomy seen earlier between Lower Bounds and Upper Bounds concerning the averaging schemes! Figure 6.14 depicts the combined (aggregate) density curve for these ten exponentials.

**The infinite chain conjecture assumes that the primary distribution as well as all intermediary distributions are defined on  $(0, +\infty)$ ,  $(-\infty, 0)$ , or  $(-\infty, +\infty)$ .** Conceptually, it might be difficult to envision how distributions drawing from both sides of the origin such as the Normal converge under chaining, but they **do** converge! The chained Normal defined on  $(-\infty, +\infty)$ , with plenty of simulated positive and negative numbers, shows a remarkable ability to converge to the logarithmic [under certain conditions to be discussed in the next chapter].

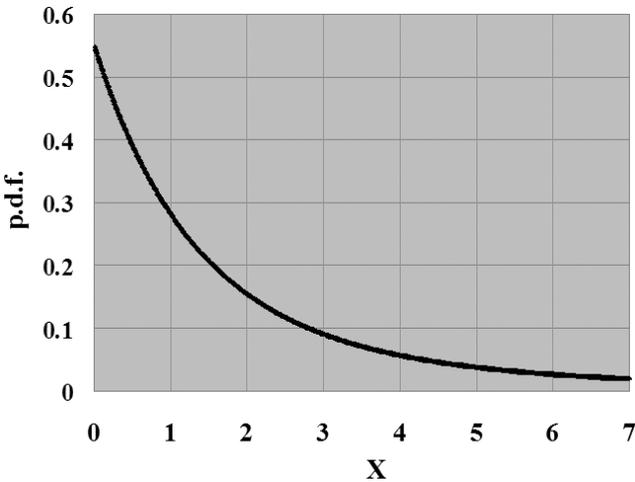


Figure 6.14 Aggregate/Combined Density of 10 Realizations of Exponential( $U(0, 1)$ )

The infinite chain conjecture can then be stated formally as follows:

Consider the following set of distributions; all having any number of parameters; **all parameters are real, not discrete; all moments/means have finite values**; all relevant parameters and all distributions are defined on such ranges as  $(0, +\infty)$ ,  $(-\infty, 0)$ , or  $(-\infty, +\infty)$ ; then, for chains made of such distributions (but limited to those types of parameters and distributions to be discussed in the next chapter) where **all parameters** are tied up to yet other distributions, and so forth, the primary distribution (the one sitting on top of the pyramid not serving as a parameter for others) is Benford in the limit as the number of sequences goes to infinity, and regardless of course of the exact values of the parameters at the very (infinitely deep) bottom of the chain.

Surprisingly, chaining distributions defined on  $(\mathbf{a}, +\infty)$  or on  $(-\infty, -\mathbf{a})$ , which do not come under the chain conjuncture, yield convergence to the logarithmic just as well, regardless of whether  $\mathbf{a}$  is an integral power of ten or not. Conceptually though, one finds it hard to explain such convergence except in cases of distributions defined on  $(+10^{\text{Integer}}, +\infty)$  or  $(-\infty, -10^{\text{Integer}})$ . Yet this behavior can be understood when distribution is not abruptly launched at  $\mathbf{a}$  and then falls off steadily, but rather gradually rises from  $\mathbf{a}$  and then falls, meaning that the portion of data near  $\mathbf{a}$  is not very large (i.e. not highly significant in influencing digital configuration).

Chains inadvertently possess some sort of 'proper' direction in order for them to converge. For example,  $U(0, U(0, U(0, M)))$  taken further to infinity (infinite number of sequences, not just three) is Flehinger-like successful scheme (with Lower Bound stuck at 0 as opposed to 1), while  $U(U(U(A, B), C), D)$  taken to infinity, where  $D > C > B > A$ , is not logarithmic at all! The former fastens LB to 0 and expand UB outward, while the latter fastens UB to the constant D and contracts LB gradually up to A. Certainly it is plausible to argue that for everyday real-life data sets, an extremely natural lower bound would be zero or one, and that distributions expand outward in the positive direction, much like the former scheme and unlike the latter.

The two-sequence uniform chain  $U(L, U(L, M))$  converges to the logarithmic not whenever difference  $M - L$  is large, but rather whenever their log difference is large! That is, whenever  $\text{LOG}_{10}M - \text{LOG}_{10}L$  is larger than roughly 8 or 9 depending on logarithmic accuracy desired. Hence  $U(0, U(0, 1))$  is by far superior to, say,  $U(1, U(1, 999))$  as there are infinitely many IPOT values on that 'short' interval between 0 and 1 (i.e. 'log distance' between 0 and 1 is infinitely long). The above discussion explains why the chain  $U(1, U(1, M))$  needs to have the value of M to be at least 10,000,000 approximately to see a good agreement with the logarithmic.

Another 'chain paradox' is found in the observation that  $U(1, U(1, 10))$  yields first digits of {36.0%, 19.6%, 14.7%, 10.1%, 7.7%, 5.3%, 3.3%, 2.6%, 0.7%} with good spread from 1 to about 6, while  $U(1, U(1, U(1, U(1, 10))))$  yields first digits of {81.0%, 11.6%, 4.3%, 1.7%, 1.0%, 0.3%, 0.1%, 0.0%, 0.0%} where most values are congregating just over 1 (the vast majority of them are below 2, and the chain does not have any meaningful spread). In fact, the longer the chain the more values are crowding out and regressing towards 1! These two opposing examples show that (under certain conditions) we could get a worsening logarithmic result with increasing number of sequences in the chain (the shorter two-sequence chain had by far better logarithmic result than that longer four-sequence one!)

Yet, when chains are fastened to 0 [as opposed to 1 as in the paradox above] the chain with more sequences (namely longer) is by far superior! That is, the longer chain of  $U(0, U(0, U(0, U(0, 1))))$  is much closer to the logarithmic than the shorter chain of  $U(0, U(0, 1))$ !

Finally, the confluence of two observations here points to a second chain conjecture. A chain-thought-experiment is performed, employing some imaginary

infinite chain of whatever form and style, and which is being called CHN. Since ultimately CHN is nothing but some particular distribution (and logarithmic as per our infinite conjecture), it is now being utilized to represent the parameter for some single-parameter distribution called Z. Clearly, the newly created chain  $Z(\text{CHN})$ , having infinite +1 sequences, is logarithmic just the same, since it is infinitely chained. Yet, a different perspective is suggested here by viewing  $Z(\text{CHN})$  merely as a two-sequence chain, and proclaiming that its logarithmic behavior springs from the fact that its parameter is being chained to a density that is logarithmic in its own right. A second similar insight leading to the same conclusion can be inferred from Benford's own attempt in establishing the logarithmic distribution by letting UB vary exponentially. His semi-logarithmic scale of P vs. UB and his calculations of the area under this curve to arrive at its average height are mathematically equivalent to the use of exponential growth series as upper bounds. His entire scheme could then be interpreted as the chain of distribution **Uniform(0, exponential series)** which indeed converges to the logarithmic. This short chain is noted for its special feature of inserting a distribution that is logarithmic in its own right, namely exponential growth series, to serve parameter **b** of the Uniform distribution. Admittedly, parameter **a** is being made fixed at 0, as opposed to being chained as well, but here it is actually beneficial not to chain **a**, only **b**. By Proposition VI which relates exponential growth series to  $k/x$  distribution, the observation made earlier in this chapter about Benford's proof, namely that the chain **Uniform(0,  $1/(x \cdot \ln 10)$  defined on (1, 10))** is indeed logarithmic, may serve as the mirror image of the above discussion.

**The second conjecture: Any two-sequence chain having all parameters of the primary density derived from logarithmic distributions is logarithmic there and then without any need to expand infinitely.**

Three chains that are very similar to Benford's own scheme have been examined via computer simulations. Results here lend strong support for the second conjecture.

- (A): **Uniform(0, Lognormal)** is progressively more logarithmic as shape parameter is gradually increased all the way from 0.1 to 0.7, culminating in a near-perfect logarithmic behavior of the chained uniform for shapes over 1.25.
- (B): **Uniform(0, exponential)** is very nearly logarithmic, but not quite so, since any exponential is only approximately so.
- (C): **Uniform(0, a chain of 15 Rayleighs)** is nearly perfectly logarithmic.

Yet, strictly (formally) speaking, none of these three examples are fully logarithmic even under the second conjecture, since none of these parametrical densities (Lognormal, exponential, 15-sequence-Rayleigh chain) is perfectly logarithmic no matter what parameter is chosen, rather they are just extremely close to it. In any case, these three simulations results certainly lend strong support for the second conjecture. The fact that in all three examples parameter  $\mathbf{a}$  is stuck at 0 does not imply any lack of support for the conjecture, because had we actually chained parameter  $\mathbf{a}$  to a distribution it would still be necessary to have it bounded by the restriction guaranteeing that  $\mathbf{a}$  is less than  $\mathbf{b}$ , as in the definition of the Uniform, namely  $\max[\text{distribution of a-parameter}] < \min[\text{distribution of b-parameter}]$ , and all this would most likely point to 0 anyhow as the value for  $\mathbf{a}$ .

Three other simulations suggestive of the second conjecture are:

**chi-sqr(the integer part of exponential growth series)** is nearly logarithmic, and regardless of rate and base, provided that the exponential series is really logarithmic.

**chi-sqr(integers of a Lognormal)** is progressively more logarithmic as shape parameter increases from 0.1 to 0.7, culminating in a near perfect logarithmic behavior for shape over 1.25.

**chi-sqr(integers of a chain of 13 uniforms - all with parameter a stuck at 0)** is nearly logarithmic.

[Note: These three results are so in spite of our earlier prohibition against using discrete parameters such as the degrees of freedom (d.o.f.) of the chi-sqr. This is an indication that the conjectures are too timidly and too cautiously stated, and that chains are applicable in many more situations. In addition, the three chains above are logarithmic in spite of d.o.f. being essentially a non-chainable shape parameter of the gamma distribution, as will be discussed in the next chapter].

The following ten chains of distributions came out quite nearly logarithmic. Since their parameters are being derived from densities that are nearly or exactly Benford, these results lend further support for the second conjecture.

**Weibull(Lognormal shape  $> 1.25$ , Lognormal shape  $> 1.25$ ) ;**

**Weibull(chain of uniforms, chain of uniforms) ;**

**Weibull( $1/(x*\ln 10*M)$  on  $(10^H, 10^{H + \text{integer}_M})$ ,  
 $1/(x*\ln 10*N)$  on  $(10^G, 10^{G + \text{integer}_N})$ ) ;**

**Rayleigh**(chain of uniforms) ;

**Rayleigh**(Lognormal shape > 1.25) ;

**Rayleigh**( $1/(x*\ln 10*N)$  over  $(10^F, 10^{F+integer\_N})$ );

**Wald**(Lognormal shape > 1.25, Lognormal shape > 1.25);

**Wald**(chain of uniforms, chain of uniforms);

**Wald**(chain of Rayleighs, chain of Rayleighs);

**Wald**( $1/(x*\ln 10*M)$  on  $(10^H, 10^{H+integer\_M})$ ,  $1/(x*\ln 10*N)$  on  $(10^G, 10^{G+integer\_N})$ ).

The general form of the second conjecture could be succinctly expressed as follows:

**AnyDensity(AnyBenford) is Benford**

Although this is true for a limited class of distributions/parameters to be discussed in the next chapter.

One could hardly argue that these are just special cases, and that what drives that nearly perfect logarithmic behavior for the above chains is the particular density form used or the specific chain arrangement, as opposed to the fact that their parameters are chained to logarithmic distributions. Such claims can be thoroughly refuted by using many convincing counter demonstrations, and in particular the demonstration in the form of the chain **Rayleigh(Lognormal)** as one example. The table in Fig. 6.15 shows simulation results of the chain Rayleigh(Lognormal), with location parameter stuck at 3, while shape is being allowed to vary (because shape is the one parameter exclusively controlling logarithmic behavior in the Lognormal). For each value of the shape parameter of the Lognormal the table gives SSD measure of the degree of logarithmic behavior of the chained Rayleigh as well as SSD for the Lognormal itself. The results here show a clear dependency of the logarithmic behavior of the chained Rayleigh on the logarithmic behavior of the Lognormal! There is a perfect correlation between their digital configurations! Here, the form of the distribution serving the parameter, namely the Lognormal, and the whole setup of the chain is a constant while the only factor that varies is Benfordness of the Lognormal, which clearly induces and controls Benfordness of the chained Rayleigh!

Lognormal Shape	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
SSD - Lognormal	1933	1682	1199	718	374	170	50	19	4	2
SSD - Chain	233	175	118	82	37	11	6	2	2	1

Figure 6.15 Shape Manipulating Digital Behavior of the Chain Rayleigh(Lognormal)

This strongly suggests in general that for the chain  $P(Q)$ , it's the Benfordness of the chainer  $Q$  that determines Benfordness of the chaineed  $P$  or, at a minimum, that Benfordness of chaineed  $P$  is not indifferent to Benfordness of chainer  $Q$ . And this also strongly suggests a daring **extrapolation of the second conjecture** by applying it in total generality also to **any** Benford or non-Benford chainer  $Q$  distribution serving as parameter for the chaineed  $P$ , and asserting that Benfordness of any chain  $P(Q)$  is at a minimum equal to the Benfordness of  $Q$ , but probably slightly higher, never less, and thus that Benfordness is always conserved in the process of chaining. In other words, that  $[\text{SSD of } P(Q)] \leq [\text{SSD of } Q]$  so that there is never a retreat or backtracking in Benfordness in the act of chaining any two distributions. One can mathematically define relative "Benfordness" not only in terms of SSD but, alternatively perhaps, in terms of how severely the digits are skewed favoring low digits over high ones, such as in Saville's Slope or ES12, and so forth, and substituting these other measures in the inequality above. The above inequality if formally proven, could serve as the setup of a rigorous mathematical proof for the first infinite conjecture (an alternative to Miller's proof), since a process generating an infinite number of tiny improvements in Benfordness should lead to an outright logarithmic behavior no matter how infinitesimal that improvement at each sequence in the chain happens to be. Such an alternative proof would probably be by far more comprehensive, covering many more cases, setups, and distribution forms; all of which are already strongly suggested by simulation results. As far as the second conjecture itself is concerned, the above inequality immediately implies it!

The results of Fig. 6.15 regarding the chain Rayleigh(Lognormal) are strictly in accordance with the extrapolation of the second conjecture, strongly confirming it. Moreover, inequality  $[\text{SSD of Rayleigh(Lognormal)}] < [\text{SSD of Lognormal}]$  here holds true without any need to utilize the equality sign.

Another computer simulation strongly in support of the inequality above and the accompanying discussion about the extrapolated second conjecture is given by the decreasing values of the sequential SSD measures of the following complex chain:

Uniform(0, Rayleigh(Rayleigh(Weibull(Uniform(0, 65), Normal(87, 5))))))  
**4.6      42.9      326.0    3977.5      377.7      5578.2**

What this five-sequence chain clearly demonstrates is that Benfordness is continuously being improved at each step in the chain, that there is never a pause or a retreat!

[SSD value of 377.7 for the Uniform(0, 65) can be overlooked, it pertains to the shape parameter of the Weibull which does not play any role in chaining, as will be shown in the next chapter].

Let us consider the following logarithmic chain:

gamma(infinite chain of uniforms, infinite chain of exponentials).

As discussed earlier, this could be viewed either as:

(I) an infinite chain, and thus logarithmic as per the first conjecture on infinite chains.  
 (II) gamma(logarithmically distributed param, logarithmically distributed param), and thus logarithmic as per the second conjecture. This flexibility in our point of view demonstrates how the second conjecture synthesizes and harmonizes nicely with the original infinite chain conjecture. Indeed this is a necessary feature to hold the whole chain edifice together and contradictions would certainly arise otherwise.

To recap what was discussed earlier: harmonious relationship and mutual dependency between the first and the second conjectures can be found employing the following argument showing how the extrapolated second conjecture explains the infinite-sequence conjecture by utilizing two assumed factors:

- [I] We have widened the scope of the second conjecture and asserted that Benfordness is always conserved in the act of chaining, that there is never a retreat or backtracking in Benfordness.
- [II] Moreover, focusing on the micro level of an infinite-sequence chain, we postulate that there is an active mechanism in most sequential jumps resulting in a dynamic tendency for the outcome to end up with some 'improvement' in Benfordness.

The confluence of these two factors above on an infinite chain is such as to insure a slow but certain drift towards complete and final Benfordness.

An immediate corollary of the (two-sequence) second conjecture is that any finite chain, however long, having all its last (lower) parameters derived from logarithmic distributions is logarithmic. This is so via repeated applications of the second conjecture from the bottom up. [The symbol **B** to be used here would signify Benfordness, i.e. a logarithmic distribution]. For example, the chain  $K(R(S(N(M(P(\mathbf{B}))))), V(H(\mathbf{B}, \mathbf{B})))$  is reduced to  $K(R(S(N(M(\mathbf{B}))))), V(\mathbf{B}))$ , then to  $K(R(S(N(\mathbf{B}))), \mathbf{B})$ , and so forth, to  $K(\mathbf{B}, \mathbf{B})$ , and finally to  $\mathbf{B}$ . In addition, the extrapolated second conjecture guarantees a quicker near-convergence for the applications of the original infinite conjecture in the case of finite chains whenever all lowest (parametrical — at the bottom) densities are themselves closer to Benford to begin with, or at least having monotonically decreasing digit distributions resembling it. In other words:

When climbing Mount Everest, if one's starting point is at a higher plateau then one reaches the summit sooner.

## CHAINABLE DISTRIBUTIONS AND PARAMETERS

---



---

In this chapter a general rule is given describing what kind of parameters or combination thereof responds favorably to chaining leading to logarithmic convergence. A detailed list showing specifically the chainability status of some classic/standard and widely used distributions is also included.

Conclusions here are obtained by way of fusing multiple empirical results together to arrive at some coherent and consistent rules. Empirical results are achieved by way of computer simulations. This involves taking classic distributions one at a time, chaining one or two of the parameters, and examining digital behavior. The second chain conjecture (of two sequences) offers an easy way of checking compliance. The setup for computer simulations is straightforward as parameters are tied either to  $1/(x*\ln10*N)$  defined over  $(10^F, 10^{F + \text{INTEGER } N})$  or to the Lognormal, turning the shape parameter there high and low as an experimental knob and observing effects. The latter choice of the Lognormal constitutes a much more convincing demonstration of any possible logarithmic convergence due to the ability to manipulate logarithmicity of the Lognormal itself via its shape parameter and to observe the effects on the chained distribution under investigation. In contrast, approximating an infinite chain by using a finite number of chained distributions in computer simulations is cumbersome and tedious, and also leads to the constant dilemma of trying to figure out the correct number of necessary sequences for sufficient accuracy. Since both conjectures are intimately connected and mirror each other, we shall assume for convenience that the result for one implies the same for the other. In other words, that the rules governing convergence for the first conjecture are identical to the rules governing convergence for the second conjecture.

### General principles:

**Scale** parameters such as  $\lambda X$  or  $X/\lambda$  (**divisions & multiplications**) in the absence of any location parameter always respond vigorously to chaining and yield the logarithmic. The requirement here is that each X in the PDF expression is

always and consistently accompanied by such  $\lambda$ , and that there isn't any other  $\lambda$  combination with X using other arithmetic operations anywhere else in the PDF expression (nor  $\lambda$  reappearing alone as addition or subtraction, such as in  $\lambda X - \lambda$ , and so forth).

**Location** parameters such as  $X - \mu$  (**subtractions**) in the absence of any scale parameter responds vigorously to chaining and yield the logarithmic wherever the range of the chained  $\mu$  is high and hovers above a certain level, depending on the particular distribution in question. Low distributed values of the chained  $\mu$  do not yield the logarithmic, yet cause it to respond to chaining a great deal, approaching the logarithmic but not quite reaching it. In other words,  $\mu$  must be chained to a distribution distributed above a certain level. For example,  $\mu$  chained to the Lognormal (0.1, 1.0) does not lead to full convergence for many densities, while  $\mu$  chained to the Lognormal (8, 1.0) leads to full convergence in most densities. This curious state of affairs is not hard to explain!

In the case of the simultaneous presence of **location and scale** parameters everywhere in the PDF expression of the form  $(X - \mu)/\lambda$ , both parameters must be chained in order to obtain the logarithmic. Again, the requirement here is that each X is always accompanied by  $\lambda$  and  $\mu$  in the same arithmetic way everywhere in the PDF expression. Chaining only one parameter while leaving the other one as a constant does not yield the logarithmic unless done in a way that yields much higher values (in comparison) for the chained parameter than the value of the other parameter being left a constant. Yet there is a sharp dichotomy here between  $\lambda$  and  $\mu$  when only one parameter is being chained; chaining only scale does not yield anything resembling the logarithmic unless done in a way that results in much higher values for scale than location. On the other hand, chaining only location always yields something quite close to the logarithmic, regardless of the relative level in values between location and scale.

When the parameters above are of the form  $\lambda*(X - \mu)$ , both parameters must be chained in order to obtain the logarithmic. Chaining only one parameter while leaving the other one as a constant does not yield the logarithmic unless done in a way that yields much higher values (in comparison) for the location parameter than the value(s) of the scale parameter. This draws a distinction between this case (where location has to be large no matter what parameter is being chained) and the previous case where  $\lambda$  appears as a denominator (where the parameter being chained has to be large). And just as in the previous case, chaining only location always yields something quite close to the logarithmic, while chaining only scale does not have any effect unless location is much larger than scale.

**Shape** parameters such as  $X^k$  (**powers**) do not yield the logarithmic when chained (that is, shape is not chainable). Interestingly, chaining only the shape of the Weibull distribution does not lead to any kind of convergence whatsoever; while the shape of the gamma responds vigorously to chaining, yet without that complete convergence. An extremely compelling explanation for this dichotomy is found in the expressions for the averages,  $\lambda^* \Gamma(1 + 1/k)$  for the Weibull and  $\lambda^* k$  for the gamma. In the case of the gamma the average is responding steadily to any increase in  $k$ , while the average of the Weibull becomes progressively more and more indifferent to further increases in  $k$  beyond a certain point since it converges to  $\lambda^* \Gamma(1)$  in the limit as  $k$  gets large. In other words, in the case of the gamma, parameter  $k$  is continuously playing a role in the determination of the average hence it's chainable to some extent, while in the case of the Weibull, parameter  $k$  does not play any role in the determination of the average beyond the initial few low values hence it's not chainable at all. The general rule for shape parameters can be then stated as follows: whenever derivative  $d(\text{average})/d(\text{parameter})$  continuously falls and then diminishes in the limit (hence parameter is deemed irrelevant to the expression of centrality in the limit) no chainability can be found whatsoever. Otherwise, when it does not diminish, it responds vigorously to chaining but without that complete convergence to the logarithmic. It is noted that the median here proves a better measure of centrality than the average in predicting and explaining chainability behavior, as might be expected perhaps for being a much more robust measure of centrality in general.

The conjecture regarding a general principle here for all parameters (not only for the shape parameter but also for scale and location parameters) can then be stated as follows: **A parameter that does not continuously involve itself in the expression of centrality is not chainable at all and does not show even feeble convergence under chaining. 'Continuously involved' means that  $d(\text{center})/d(\text{parameter})$  never diminishes as parameter gets large.** The converse is not always being exactly observed, as there are cases where parameters that do involve themselves in an additive expression of centrality turned out not to be chainable at all (suggesting perhaps that additive expression of centrality is weaker than multiplicative expression). Three such cases come to mind: the generalized exponential  $1/\lambda^* \exp(-(X-\mu)/\lambda)$ , Fisher-Tippett, and the Normal distributions. Their medians are respectively:  $1.000^* \mu + 0.693^* \lambda$ ,  $1.000^* \mu + 0.367^* \lambda$ , and  $1.000^* \mu + 0.000^* \sigma$ , emphasizing (with a factor of 1) the location parameter  $\mu$  which responds vigorously but not completely to chaining

[being the only parameter chained], and de-emphasizing (with factors less than 1) scale which is in fact totally indifferent to chaining [being the only parameter chained]. Again, this example shows that the more a given parameter is involved in the expression of centrality (higher value of coefficient such as 1) the more it contributes to chainability. Yet, the more appropriate view to take here is to observe the simple fact that basically (in the first two distributions above) both  $\mu$  and  $\lambda$  are involved in the expression of centrality, hence ideally both should be chained for a full and complete convergence to be observed, not just one of them. Cases where the two parameters are involved in a multiplicative expression of centrality (such as  $\mu*\lambda$ ) seem to lend each single parameter sufficient weight as to be chainable or almost so in and by itself.

**Why should only those parameters that are strongly and continuously involved in the expression of centrality be chainable? Conceptually the answer is quite straightforward: because the mere act of chaining such parameters in any way at a minimum yields a set of distributions with increasing (stretching) upper bounds, derived from the fact that that parameter affects location (centrality) a great deal, and hopefully with lower bounds staying fixed near 0, or 1, or any other low number (preferably IPOT). As was seen in the case of the simple averaging schemes, such a setup results in an overall density that is diminishing, having that one-sided tail to the right so typical of logarithmic distributions. Nothing of the sort could happen if the parameter to be varied by chaining does not affect centrality in the least!**

Let us recall the short chain  $exponential(Uniform(0, 1))$  in Figs. 6.13 and 6.14. The parameter being chained in this case is the scale  $\rho$ , which is also exactly the expression for the average! Therefore, varying  $\rho$  there by way of chaining it to the Uniform implies that we are simply stretching upper bounds, while lower bounds are fixed at 0 by default, that is, by virtue of it being the exponential distribution. Visualization of this process supports the conceptual argument given here, and it is quite exuberating to find such harmonious and consistent results in the relation between a wide variety of computer simulations of chains and the general understanding gained in the study of averaging schemes where lower/upper bounds interplay strongly influences logarithmic behavior.

The second conjecture makes the mechanization at play here becomes a bit more apparent. If  $exponential(Benford)$  comes with logarithmically-stretched upper bounds where the small is numerous and the big is rare, then concentration of the

exponentials themselves on the left is even higher, falling off much more rapidly on the right, and therefore perhaps outright logarithmic behavior can be expected for the chain itself.

**In cases where the scale parameter  $\lambda$  and  $X$  are of different dimensions, no chainability can be obtained. Hence  $X/\log(\lambda)$ ,  $X/\ln(\lambda)$ ,  $X/\sqrt{\lambda}$ ,  $X/(\text{Nth root of } \lambda)$  in the PDF expression do not lend themselves to chainability.** On the other hand, powers of  $\lambda$  in the PDF such as  $X/\lambda^2$ ,  $X/\lambda^{3.5}$ , etc. lead to chainability given that power is 1 or higher. This is so because any power transformation (of 1 or higher) of any logarithmic data or variable (or parameter!) is also logarithmic, hence chaining  $\lambda$  to a logarithmic distribution implies that  $\lambda^2$  or  $\lambda^{3.5}$  are also logarithmic, and thus could be considered as a whole the actual parameter serving the distribution instead of  $\lambda$ , and as such having the same dimension as  $X$ . On the other hand, log transformation, square root transformation, or any Nth root transformations of data do not result in any logarithmic inheritance of data, hence they are not chainable. The same pessimistic argument applies to  $\log(X)/\lambda$ ,  $(\sqrt{X})/\lambda$ , Nth root( $X$ )/ $\lambda$  situations where  $X$  is of a different dimension than that of  $\lambda$ .

The two tables in Figs. 6.16 and 6.17 summarize many important simulation results. The designation “**BEN**” indicates an outright Benford behavior for the chain, namely chainability; “**NOT**” indicates a total indifference to chaining, not even having a feeble improvement towards the logarithmic in digital configurations; while “**some**” in lower case letters indicates a lack of full logarithmic behavior yet having some definite and vigorous response to chaining resulting in a considerable movement towards the logarithmic distribution. All chain results in these two tables are obtained by tying each parameter to a Lognormal distribution, turning the shape parameter high and low as an experimental knob, and observing the variations of digital configuration of the distribution under investigation to determine how responsive it is to the Benfordness of its distributed parameters.

#### Comments:

For the single parameter distributions of the **exponential** and **uniform** on  $(0, b)$ , the parameter is explicitly present in the expression for the center, hence full chainability.

The **origin-center Normal** ( $\mu = 0$ , and where s.d. is being chained) should be viewed as the combination of two separate distributions, one on  $(-\infty, 0)$  and the other on  $(0, +\infty)$ , each with its own center expressed explicitly in terms of s.d. parameter  $\sigma$ , hence chainability for each, which in turn implies chainability for the combination. This result is in spite of the fact that the overall mean

Distribution	pdf expression	Param 1	Param 2	both 1,2	MEAN
Weibull (0, +∞)	$(k/\lambda)(x/\lambda)^{(k-1)}e^{-(x/\lambda)^k}$	λ BEN	k NOT	λ & k BEN	λ*Γ(1+1/k)
Rayleigh (0, +∞)	$x \exp\left(\frac{-x^2}{2\sigma^2}\right) / \sigma^2$	BEN			$\sigma\sqrt{\frac{\pi}{2}}$
Exp 1 (0, +∞)	$\rho^* \exp(-\rho x)$	BEN			1/ρ
Exp 2 (0, +∞)	$(1/\rho)^* \exp(-x/\rho)$	BEN			ρ
Exp 3 (0, +∞)	$(1/\sqrt{\rho})^* \exp(-x/\sqrt{\rho})$	some			$\sqrt{\rho}$
Exp 4 (0, +∞)	$(1/\rho^{7.5})^* \exp(-x/\rho^{7.5})$	BEN			ρ <sup>7.5</sup>
Exp 5 (0, +∞)	$(1/\rho^8)^* \exp(-x/\rho^8)$	BEN			ρ <sup>8</sup>
Exp 6 (0, +∞)	$(1/\log(\rho))^* \exp(-x/\log(\rho))$	some			log(ρ)
Gamma (0, +∞)	$x^{k-1} \exp(-x/\theta) / \Gamma(k) \theta^k$	θ BEN	k some	θ & k BEN	θk
Wald (0, +∞)	$\left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp\left(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right)$	μ BEN	λ NOT	μ & λ BEN	μ
LogNormal (0, +∞)	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{[\ln(x)-\mu]^2}{2\sigma^2}\right)$	σ NOT	μ NOT	σ & μ NOT	exp(μ+σ <sup>2</sup> /2)
Uniform a=0 (0, b) or (0, +∞)	1/b	BEN			b/2
Gompertz (0, +∞)	$b e^{-bx} e^{-\eta e^{-bx}} [1 + \eta (1 - e^{-bx})]$	b BEN	η some	b & η BEN	long & complex expression
Nakagami (0, +∞)	$\frac{2\mu^\mu}{\Gamma(\mu)\omega^\mu} x^{2\mu-1} \exp\left(-\frac{\mu}{\omega} x^2\right)$	ω some	μ NOT	ω & μ some	$\frac{\Gamma(\mu + \frac{1}{2})}{\Gamma(\mu)} \left(\frac{\omega}{\mu}\right)^{1/2}$
Gupta Kundu (0, +∞)	$\alpha\lambda^* \exp(-\lambda x) * (1 - \exp(-\lambda x))^{(\alpha-1)}$	λ BEN	α NOT	λ & α BEN	1/λ * [ψ(α+1) - ψ(1)] ψ being the digamma

Figure 6.16 Chainability of Some Distributions Defined on (0, +∞)

here is a constant zero no matter what value parameter σ takes, and so it may be erroneously claimed that the mean 0 is indifferent to parameter σ! This apparent contradiction is misguided because the correct view to take here is that each side yields its own separate mean as a function of parameter σ (both of opposite sign and of the same absolute value) and only the aggregate mean is zero.

For the **Wald**, the expression for the center involves only location parameter μ, omitting scale λ altogether, hence μ is fully chainable while λ is totally indifferent to chaining. If one wishes to view this dichotomy here between μ and λ in terms of the conjecture that d(center)/d(parameter) must not diminish for chainable parameters, then one could simply observe [using partial derivative notations and interpretation] that since mean(μ, λ) = μ, therefore ∂(mean)/∂(μ) = 1, meaning

Distribution	pdf expression	Param 1	Param 2	both 1,2	MEAN
Normal ( $-\infty, +\infty$ )	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\sigma$ NOT	$\mu$ some	$\sigma$ & $\mu$ BEN	$\mu$
Origin-centered Normal ( $-\infty, +\infty$ )	$1/\sigma\sqrt{(2\pi)} * \exp(-x^2/2\sigma^2)$	BEN			0
Fisher-Tippett ( $-\infty, +\infty$ )	$(1/b)*e^{-(x-\mu)/\lambda} * \exp(-e^{-(x-\mu)/\lambda})$	$\lambda$ NOT	$\mu$ some	$\lambda$ & $\mu$ BEN	$\mu + 0.57721*\lambda$
Logistic ( $-\infty, +\infty$ )	$\frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2}$	$s$ NOT	$\mu$ some	$s$ & $\mu$ BEN	$\mu$
Cauchy-Lorentz ( $-\infty, +\infty$ )	$\frac{1}{\pi\gamma \left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}$	$\gamma$ NOT	$X_D$ BEN (almost)	$\gamma$ & $X_D$ BEN	infinite and hence not part of our conjecture
Pareto [ $a, +\infty$ )	$(\theta/a)(x/a)^{-(\theta+1)}$	$a$ BEN	$\theta$ NOT	$a$ & $\theta$ BEN	$a^{*\theta/(\theta-1)}$ if $\theta > 1$ $\infty$ if $\theta < 1$
Generalized exp1 [ $\mu, +\infty$ )	$\rho * \exp(-\rho * (x-\mu))$	$\rho$ NOT	$\mu$ some	$\rho$ & $\mu$ BEN	$\mu + 1/\rho$
Generalized exp2 [ $\mu, +\infty$ )	$(1/\rho) * \exp(-(x-\mu)/\rho)$	$\rho$ NOT	$\mu$ some	$\rho$ & $\mu$ BEN	$\mu + \rho$

Figure 6.17 Chainability of Some Distributions Defined on  $(-\infty, +\infty)$  &  $(\mu, +\infty)$

that  $\mu$  is **always** involved in the expression of the mean, while  $\partial(\text{mean})/\partial(\lambda) = 0$ , meaning that  $\lambda$  is **never** involved in the expression of the mean.

On the face of it, the case of the **Lognormal** may appear quite puzzling because neither parameter is chainable in any way (not even when both are chained simultaneously) even though both are decisively involved in the expression of the mean, and at least  $\mu$  is also involved in the expression of the median. But the absence of any chainability in the case of the Lognormal springs from that severe dimensional mismatch there (between  $X$  and  $\sigma$ , and between  $X$  and  $\mu$ ), as it uses  $(\ln(X) - \mu)/\sigma$  in its PDF expression. Surely whenever shape parameter of the Lognormal is sufficiently high it is almost perfectly logarithmic there and then without any need to chain anything, although the concern here is to obtain the logarithmic via chaining whenever shape value is low.

The **Nakagami** distribution suffers from a lack of dimensional compatibility between  $X$  and both parameters  $\mu$  and  $\omega$ , hence not chainable.

These last two cases, the Lognormal and the Nakagami, are the only distributions found where even though all parameters are chained, it's all futile and no convergence is seen. They represent the only glaring and stubborn 'exceptions' to the chain idea.

For the **Gupta Kundu** distribution, shape parameter  $\alpha$  is present in the expression for the mean  $(1/\lambda)*[\psi(\alpha+1) - \psi(1)]$ , however, the expression inside the brackets quickly converges to a number slightly less than 6 and  $d(\text{mean})/d(\text{parameter}-\alpha)$  quickly diminishes, hence its total lack of chainability. The same reasoning applies to the expression for the median there, which is  $(-1/\lambda)*\ln[1 - 1/(\alpha \text{ th root of } 2)]$ . Parameter  $\lambda$  on the other hand is continuously involved in the expressions of centrality, hence its chainability.

The mean and the median for the **Pareto** distribution are  $a*\theta/(\theta-1)$  and  $a*(\theta \text{ th root of } 2)$ , respectively, whenever  $\theta \geq 1$ . Hence  $\theta$  becomes progressively less relevant in the expression of centrality and as a consequence it is not chainable at all. Parameter  $a$ , on the other hand, is continuously involved in the expression of centrality and leads decisively to the logarithmic under chaining as expected.

The center of the **Normal** is  $\mu$  [the mean], hence chaining  $\mu$  alone leads to strong logarithmic tendency but not fully. Chaining only  $\sigma$  [the s.d.] does not lead to any logarithmic convergence [unless  $\sigma \gg \mu$ , a situation where  $\sigma$  now controls overall location and spread on the x-axis much more than  $\mu$  does, akin to the origin-center Normal with  $\mu = 0$ ]. In any case, the overall correct view here is to require both parameters  $\sigma$  and  $\mu$  to be chained regardless of which is involved in the expression of centrality more and which is involved less because, in extreme generality, both are involved in the determination of centrality and the spread of values on the x-axis, hence chaining both leads to Benford.

The first and the second chain conjectures were given a rigorous mathematical proof by Steven Miller covering a wide range of many classes of distributions/parameters. Computer simulations and conceptual reasoning though clearly point to further applicability, by far more so than in Miller's relatively restricted cases. The author is indebted to Miller for endowing mathematical respectability to these results which were previously backed only by Monte Carlo computer simulations. Miller then further proposed his own interpretation of the chain as a 'product of independent distributions', a vista which the author could not at all agree with. Conceptually such an interpretation is extremely hard to defend, and parameters must be assigned arbitrary values in order for any computer algorithm of such a product to be defined and implemented. Moreover, such an interpretation is

impossible to make when dealing with distributions having more than one parameter, such as the Normal, Weibull, Gamma, and so forth. Each distribution comes with a strict hierarchy of relative frequency and relative position in the pyramid, and this can never be reflected in a simple multiplicative arrangement. Moreover, the MCLT implies a symmetrical related LOG density (i.e. the Normal) for multiplications of random variables, while related LOG densities of chains are empirically found to be decisively asymmetrical. In recent communication with Miller he has fully retracted his interpretation — to the great relief of the author.

As a final comment about the chains, it is noted that for the infinite chain scheme, particular or arbitrary values used to conduct them (the 'last' numerical parameters for the independent distributions at the 'bottom') become less relevant as we consider longer and longer chains [that is, in actual 'very finite' Monte Carlo computer simulations]. An infinite chain would completely and thoroughly remove any supposed dependency on those 'last' values whatsoever, while perfectly converging to the logarithmic at the same time. In other words, as we progress in the construction of that infinite chain, improvement in logarithmic results and independence on parametrical values go hand in hand.

## FRANK BENFORD'S AVERAGING SCHEME AS A DISTRIBUTION CHAIN

---

Benford's own attempt in his celebrated 1938 article to explain the logarithmic distribution can be viewed in terms of the second chain conjecture. Benford employed a modified version of the simple averaging scheme to successfully arrive at the logarithmic distribution. His model introduces a curious twist, putting more emphasis on shorter intervals than longer ones by having Upper Bounds (UB) vary exponentially. In other words, instead of having UB vary upwards steadily by constantly adding one integer at a time, as have been done with all the averaging schemes including that of Flehinger, Benford varied it exponentially. For example, instead of letting UB vary as in  $\{99, 100, 101, \dots, 997, 998, 999\}$  and so forth, Benford let them vary as in 2% exponential growth  $\{99.0, 101.0, 103.0, 105.1, 107.2, \dots, 927.8, 946.4, 965.3, 984.6\}$  and fractions ignored. His semi-logarithmic scale of P versus UB and his calculations of the area under this curve to arrive at its average height is equivalent to the use of exponential growth series as upper bounds.

His scheme could then be interpreted simply as the two-sequence chain of distribution **Uniform(0, INT(*exponential growth*))** where the INT function yields the integral part of its argument. This very short chain is noted for its special feature of inserting a distribution that is (almost) logarithmic in its own right to serve parameter **b** of the Uniform(a, b) distribution, while parameter **a** is fixed at 0.

Benford's construction is an algorithm that blends the random and the deterministic. Let us recap Benford's model and assign specific values for the model. Lower bounds for all intervals are fixed at the low value of 1. Upper bounds vary as an exponential with 2% growth, starting from 99 and ending just short of 999. For the sake of facilitating calculations, only the integers are to be considered on each line, without effecting result significantly, hence upper bounds are  $\{99, 101, 103, 105, 107, \dots, 928, 946, 965, 985\}$ . Figure 6.18 illustrates the model.

Steven Miller's rigorous mathematical proof of the second chain conjecture in some restricted cases does cover Benford's model, but an alternative and simplified proof shall be presented here for this particular case. We have depicted in Fig. 6.18

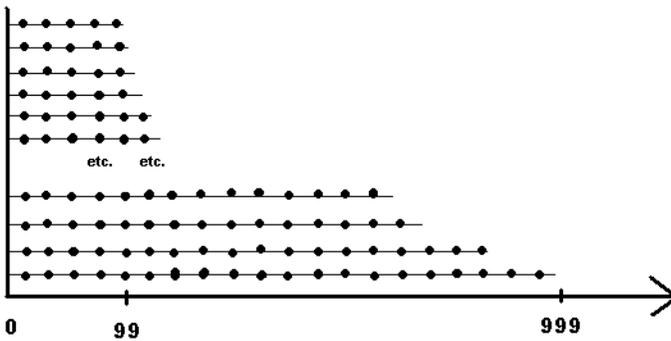


Figure 6.18 Frank Benford's Model with Upper Bounds Growing Exponentially

as dots the integers which are placed on the continuous line and showed only a fraction of them for lack of space. A line here corresponds to what was called an 'interval' in the averaging schemes. As in our earlier effort to represent the simple averaging scheme as a single data set [Fig. 4.56], it is understood that dots on shorter lines are to be repeated, and that the shorter the line the more such duplications are needed, until all lines are of the common length 999. This is so since all lines/intervals are accorded equal importance in the overall scheme, as seen in the act of taking the simple average of the digit distributions of all lines/intervals. The factor of duplication for each line is given by  $F = 999/UB$ , where  $UB$  represent the value of  $x$  at end of the line/interval. Thus the factor is simply  $F = 999/x$ . This factor  $F$  insures that all lines/intervals are now accorded equal importance (equal length). Hence the contribution of each line in generating dots (integers) for the whole scheme is the number of dots it carries up to its upper bound times the factor  $F$ , and this total per line is the same for all of them, namely 999 dots. An equivalent mirror image of this scheme is not to stretch out the lines, but instead to envision all the lines as having different densities of dots, with shorter ones denser, longer ones less dense, and all ending up with equal total numbers of dots due to this exact trade-off. The expression for the density (dots per unit length) for each line is then  $[UB * F]/UB$ , namely  $[UB * (999/UB)]/UB$ , or simply  $999/UB$ . For example, the shortest dense line here with  $UB = 99$  has density of dots  $999/99$  or about 10. The longest diluted line with  $UB = 999$  has density of  $999/999$  or just 1.

As we move in the positive direction on the  $x$ -axis anywhere between 99 and 999, the number of **lines per unit length** are falling off as in the exponential series, namely at the rate of  $-k/x$ . This is so since discrete exponential growth

series corresponds to the continuous  $k/x$  distribution via their common related uniform log as seen earlier in Proposition VI. Since each line carries a density of  $999/UB$  **line's dots per unit length** or simply  $999/x$ , it follows that the decrease in the number of **total dots per unit length** is expressed as:  $-k/x$  times  $999/x$ , and we get:  $d(\text{dots})/dx = (-k/x)*(999/x) = -Q/x^2$ , with  $Q$  being some positive constant. Anti differentiating we get: dots per unit length =  $Q/x$ . We have thus established that the density is of the type  $k/x$ . Since this is considered over  $(99,999)$  or roughly  $(10^2, 10^3)$ , exponent difference is an integral, the range is bounded by two integral powers of ten, and thus by Corollary I it is logarithmic.

Benford's scheme is then ultimately equivalent to the two-sequence chain of distributions: **Uniform(0,  $1/(x*\ln 10)$  over (100, 1000))**.

Had Benford attempted to argue in favor of the logarithmic by way of his model because it is based on the number line  $X$  itself as a general proof, then a counter argument could be raised that then the law should be universal, that is, true for all data sets, yet in fact there are many instances where real-life data is not logarithmic at all. More importantly though, Benford's scheme cannot constitute an explanation for the phenomena in general even though his abstract model is exactly logarithmic, because it would be very hard to argue that typical everyday data correspond to the structure of his model. The crux of the matter in each attempt at explaining the law is to show a strong correspondence between the abstract mathematical model and the data type in question. For example, Hill's model is perfectly suited for being a good representative of a large random pick of numbers from a wide variety of newspaper and magazines, or for the large sample of 34,269 values obtained from a mixture of 70 different topics to be discussed in Chapter 110. Yet Hill's model is not suited at all for, say, earthquake data, population data, accounting data, address data, and others, to mention just a few cases.

## EFFECTS OF PARAMETRICAL TRANSFORMATIONS ON LEADING DIGITS

---

Digital configuration of a given statistical distribution clearly depends on assigned parametrical values. For example, computer simulations for the distribution Normal (mean, s.d.) give the following results for the first digits based on different parametrical combinations:

Normal(77, 1) — { 0.0, 0.0, 0.0, 0.0, 0.0, 7.7, 63.7, 28.1, 0.5}  
 Normal(60, 20) — { 4.0, 4.3, 9.3, 15.0, 19.2, 19.3, 14.6, 9.6, 4.7}  
 Normal(30, 10) — {13.8, 35.7, 32.9, 13.8, 2.3, 0.3, 0.2, 0.5, 0.5}  
 Normal(15, 5) — {68.5, 16.3, 0.5, 0.7, 1.4, 1.9, 2.5, 3.6, 4.5}

These results should seem compatible with the overall range of where the bulk of the data lies, all of which is directly derived from the values of the parameters. For example, the bulk of the data in Normal(77, 1) falls around 77, hence digit 7 takes a strong lead there. The bulk of the data in Normal(15, 5) falls between 0 and 30, hence digit 2 and especially digit 1 obtain most of the leadership [by Chebyshev's theorem, 8/9 of the distribution is within 3 s.d. of the mean, hence approximately 89% of the distribution falls over a range of  $([15 - 3*5], [15 + 3*5])$ , i.e. over (0, 30)].

The Rayleigh( $\sigma$ ) distribution involves a single scale parameter  $\sigma$ , its mean is  $\sigma*\sqrt{(\pi/2)}$ , and its standard deviation is  $\sigma*\sqrt{((4 - \pi)/2)}$ . The computer-simulated digital results displayed below should seem compatible in general with the implied range of where the bulk of the data falls.

Rayleigh(1) — {49.3, 14.4, 4.1, 4.2, 4.9, 5.3, 5.7, 6.6, 5.5}  $m = 1.25$ ,  $sd = 0.7$   
 Rayleigh(2) — {28.4, 28.3, 20.0, 10.4, 4.8, 2.5, 1.6, 1.8, 2.1}  $m = 2.51$ ,  $sd = 1.3$   
 Rayleigh(3) — {15.2, 19.0, 19.4, 16.8, 12.2, 8.2, 4.6, 3.0, 1.6}  $m = 3.76$ ,  $sd = 2.0$   
 Rayleigh(5) — {19.3, 10.3, 11.3, 11.8, 11.8, 11.4, 9.1, 8.3, 6.7}  $m = 6.27$ ,  $sd = 3.3$

As the focus of the mean shifts from 1.25 to 6.27, so does the bulk of digital leadership!

For all classic probability distributions of the form  $(1/\lambda)*f(X/\lambda)$  and  $(1/\lambda)*f((X - \mu)/\lambda)$ , the simultaneous transformation of the  $\lambda$  scale and  $\mu$  location parameters by the same multiplicative factor  $F$  is equivalent to multiplying each value of the distribution-generated data set by  $F$  (i.e. rescaling it by  $F$ ). Whether shape parameter exists or not is irrelevant in this context, but if a shape parameter is present it needs to stay frozen (constant) in all the discussions of this chapter. As an example, in two of the four Rayleigh cases above, Rayleigh(1)  $\rightarrow$  Rayleigh(5) amounts to multiplying the single scale parameter by  $F = 5$ , which in turn implies that the distribution-generated data set increases fivefold, hence the mean advances accordingly **1.25  $\rightarrow$  6.27** ( $5*1.25 \approx 6.27$ ).

To emphasize once again that simultaneous parametrical transformations (of scale and location, regardless of shape) are nothing but a scale change, **7\*Normal(5, 2)** — namely the data set generated when each realized value from the Normal(5, 2) is multiplied by the value 7 — is perfectly equivalent to the data set of realized values from the **Normal(35, 14)**.

In the context of Leading Digits, whenever  $F$  happens to be an IPOT number [such as 10, 100, 1000, and so forth], the effect on digital configuration is totally neutral since distribution-generated data advances exactly by IPOT number and therefore mantissa and digits for each value remain unchanged. For example, transformations such as  $4.7 \rightarrow 47$  or  $3.91 \rightarrow 391$  are digit-invariant. For distributions with scale parameter of the forms  $(1/\lambda)*f(X/\lambda)$  in their probability density function (without location parameter), such as the exponential, Gamma, Weibull, Rayleigh, and others, transformation of the scale parameter via multiplication by an IPOT value [while shape is frozen] leaves digit distribution unaffected regardless of whether the distribution is logarithmic or not. For distributions with scale and location parameters of the form  $(1/\lambda)*f((X - \mu)/\lambda)$  such as the Normal, Fisher-Tippett, Logistic, Cauchy-Lorentz, and others defined over  $(-\infty, +\infty)$ ,  $(0, +\infty)$ , or  $(-\infty, 0)$ , only the simultaneous transformations of both parameters using the same IPOT value [while shape parameter stays frozen] yield such digital invariance.

The following results for the Normal and Rayleigh distributions applying simultaneous parametrical transformations by IPOT numbers confirm the above

discussion as digital configuration are invariant under such rescaling [the slight deviations here are due to the fact that results are obtained via computer simulations, not by way of exact integral calculations for areas under the curve]:

Rayleigh(0.07) — {38.0, 6.8, 6.0, 7.1, 8.2, 8.8, 9.0, 8.0, 8.1}  $m = 0.09$ ,  $sd = 0.05$

Rayleigh(0.70) — {38.5, 6.5, 6.4, 8.2, 8.4, 7.1, 8.9, 8.3, 7.7}  $m = 0.88$ ,  $sd = 0.46$

Rayleigh(7.00) — {38.1, 6.5, 6.6, 7.0, 8.5, 8.1, 9.0, 8.2, 8.0}  $m = 8.77$ ,  $sd = 4.59$

Rayleigh(70.0) — {38.3, 6.1, 6.1, 7.5, 8.0, 8.9, 8.7, 8.4, 8.0}  $m = 87.7$ ,  $sd = 45.9$

Normal(0.9, 0.1) — {15.9, 0.0, 0.0, 0.0, 0.1, 2.3, 13.5, 34.2, 34.0}

Normal(9, 1) — {15.5, 0.0, 0.0, 0.0, 0.1, 2.2, 14.0, 33.8, 34.4}

Normal(90, 10) — {15.6, 0.0, 0.0, 0.0, 0.1, 2.3, 14.1, 33.7, 34.2}

Normal(900, 100) — {15.8, 0.0, 0.0, 0.0, 0.2, 2.3, 13.6, 35.2, 33.0}

These parametrical transformations should not be confused with the scale invariance principle where data itself is being directly transformed, while here we actively transform the parameters, which only indirectly transforms the variable itself. Moreover, the scale invariance principle applies only to logarithmic distributions, while here we prove invariance in total generality where digital configuration (logarithmic or otherwise) is shown to be indifferent to any multiplicative IPOT parametrical transformation [while no such invariance claim is being stated for non-IPOT parametrical transformations]. To recap, the scale invariance principle applies to any factor, while here only IPOT parametrical transformations are considered and shown to be invariant for all types of distributions, logarithmic as well as non-logarithmic. This fact is not just about Benford's Law, but about Leading Digits in general, namely that the digital signature of the distribution is invariant under such IPOT-multiplicative transformations of the parameters. Certainly all this can be extrapolated to any base in other number systems where integral powers of the base are used as factors transforming the parameters.

Let us prove that for distributions such as  $PDF(X) = (1/\lambda)*f((X - \mu)/\lambda)$ , simultaneous multiplicative transformations of  $\lambda$  scale and  $\mu$  location parameters by the same factor  $F$  amounts simply to a multiplicative transformation of the distribution-generated data by  $F$  (i.e. rescaling data by  $F$ ). Define  $Y = FX$ , hence by The Transformation Technique [Freund's book "Mathematical Statistics", 6th Edition, Chapter 7.3, Theorem 7.1]  $PDF(Y) = (1/\lambda)*f((Y/F - \mu)/\lambda)*|1/F| = (1/(F\lambda))*f((F/F)*(Y/F - \mu)/\lambda) = (1/(F\lambda))*f((Y - F\mu)/F\lambda)$ . Hence PDF of  $FX$  is

identical to the PDF of X except that both parameters have been multiplicatively transformed by the same factor F, namely:  $\lambda \rightarrow F\lambda$  and  $\mu \rightarrow F\mu$ . An immediate corollary is that whenever  $F = \text{IPOT}$  such transformation is digitally invariant. [Note: F is assumed to be positive, that is  $F > 0$ ].

A formal proof in the first digit sense shall now be given for this invariance principle of the simultaneous multiplicative transformations of scale and location parameters by the same IPOT value. Although this proof may be considered trivial, since the above proof already demonstrates that such parametrical transformation amounts simply to rescaling by IPOT, which in turn leads to digital invariance, yet the format of the following proof is the typical starting point of many proofs in the literature regarding Benford's Law, and therefore it is quite instructive in and of itself.

Assume a density function defined on interval  $(-\infty, +\infty)$  of the form  $\text{PDF}(x) = \frac{1}{b} * f\left(\frac{x-a}{b}\right)$  where each x appears as  $\left(\frac{x-a}{b}\right)$  everywhere in the PDF expression. We divide the entire interval of  $(-\infty, +\infty)$  into  $(0, +\infty)$  &  $(-\infty, 0)$ , and in that order for the next definite integrals evaluations.

$$\text{Prob}(\text{1st digit} = d | a_0, b_0) =$$

$$\sum_{\text{int}=-\infty}^{\text{int}=\infty} \int_{(d)*10^{\text{int}}}^{(d+1)*10^{\text{int}}} \left(\frac{1}{b_0}\right) * f\left(\frac{x-a_0}{b_0}\right) dx +$$

$$\sum_{\text{int}=-\infty}^{\text{int}=\infty} \int_{-(d+1)*10^{\text{int}}}^{-(d)*10^{\text{int}}} \left(\frac{1}{b_0}\right) * f\left(\frac{x-a_0}{b_0}\right) dx$$

Where **int** runs through all the integers.

Let us now examine first-digits distribution for another density of the same form where both  $a_0$  and  $b_0$  are being simultaneously multiplied by the same factor  $10^M$ , M being any integer.

$$\sum_{\text{int}=-\infty}^{\text{int}=\infty} \int_{(d)*10^{\text{int}}}^{(d+1)*10^{\text{int}}} \left(\frac{1}{10^M b_0}\right) * f\left(\frac{x-10^M a_0}{10^M b_0}\right) dx +$$

$$\sum_{\text{int}=-\infty}^{\text{int}=\infty} \int_{-(d+1)*10^{\text{int}}}^{-(d)*10^{\text{int}}} \left(\frac{1}{10^M b_0}\right) * f\left(\frac{x-10^M a_0}{10^M b_0}\right) dx$$

A change of variables  $u = x/10^M$  or equivalently  $x = u * 10^M$  implies the relationship  $dx/du = 10^M$  and yields:

$$\sum_{int=-\infty}^{int=+\infty} \int_{(d)*10^{int}/10^M}^{(d+1)*10^{int}/10^M} \left( \frac{1}{10^M b_0} \right) * f \left( \frac{u10^M - 10^M a_0}{10^M b_0} \right) 10^M du +$$

$$\sum_{int=-\infty}^{int=+\infty} \int_{-(d+1)*10^{int}/10^M}^{-(d)*10^{int}/10^M} \left( \frac{1}{10^M b_0} \right) * f \left( \frac{u10^M - 10^M a_0}{10^M b_0} \right) 10^M du$$

The term  $10^M$  nicely cancels out twice in the PDF expression, and integral reduces to:

$$\sum_{int=-\infty}^{int=+\infty} \int_{(d)*10^{[int-M]}}^{(d+1)*10^{[int-M]}} \left( \frac{1}{b_0} \right) * f \left( \frac{u - a_0}{b_0} \right) du +$$

$$\sum_{int=-\infty}^{int=+\infty} \int_{-(d+1)*10^{[int-M]}}^{-(d)*10^{[int-M]}} \left( \frac{1}{b_0} \right) * f \left( \frac{u - a_0}{b_0} \right) du$$

Since index **int** runs over all possible integers including negative ones, the introduction of integral M value into integration limits does not have any effect, and so M can be omitted. We are left with:

$$\sum_{int=-\infty}^{int=+\infty} \int_{(d)*10^{int}}^{(d+1)*10^{int}} \left( \frac{1}{b_0} \right) * f \left( \frac{u - a_0}{b_0} \right) du +$$

$$\sum_{int=-\infty}^{int=+\infty} \int_{-(d+1)*10^{int}}^{-(d)*10^{int}} \left( \frac{1}{b_0} \right) * f \left( \frac{u - a_0}{b_0} \right) du$$

Which yields the same expression for the first-leading-digits probability of the original PDF =  $\left( \frac{1}{b_0} \right) * f \left( \frac{x-a_0}{b_0} \right)$  and which completes the proof. Certainly, this proof does not claim that simultaneously transforming location and scale parameters by the same IPOT values does not alter the distribution itself, as indeed it does, and specifically it re-scales the distribution exactly by that IPOT number. It is only with regards to Leading Digits that invariance is found here.

## DIGITS OF THE WALD, WEIBULL, CHI-SQR, AND GAMMA DISTRIBUTIONS

---

---

**For the Wald distribution**, the higher the location parameter  $\mu$  ( $mu$ ) the closer to the logarithmic the distribution is for a given value of the  $\lambda$  ( $lambda$ ) parameter. The lower the  $lambda$  the closer to the logarithmic it is for a given  $mu$  value. Hence for logarithmic behavior to occur here it is necessary to obtain a low  $\lambda/\mu$  ( $lambda/mu$ ) ratio. Roughly speaking, whenever  $\lambda/\mu$  is lower than 0.85 near-perfect logarithmic behavior is found.

**The Weibull distribution** is nearly perfectly logarithmic whenever the shape parameter  $k$  is very low — roughly whenever it's 0.7 or lower — regardless of what value the scale  $\lambda$  ( $lambda$ ) parameter takes.

**The chi-sqr distribution** is nearly logarithmic whenever the value of the degrees of freedom (d.o.f.) is 1, approximately logarithmic when it's 2, and becomes progressively less and less logarithmic for higher d.o.f. parametrical values.

**The gamma distribution** is nearly logarithmic whenever the shape parameter  $\alpha$  is approximately lower than 0.7.

## DIGITAL PATTERNS OF THE EXPONENTIAL DISTRIBUTION

---

---

Leading digits of the exponential distribution are never perfectly logarithmic regardless of the value the parameter takes, but they are fairly close to it. Fascinating repeated patterns are seen whenever digit distributions are (vertically) plotted versus parameter  $p$  (horizontally), as if [digits are] being considered as a variable, namely as a function of parameter. Individual digits nicely oscillate consistently and devotedly around their central logarithmic values of  $\text{LOG}(1 + 1/d)$  indefinitely in ever-widening cycles as parameter  $p$  increases. Unfortunately, the nine cycles of the nine digits are not synchronized. Rather, they are continuously out of step with each other, making it completely impossible to find one particular parameter able to simultaneously seize all nine digits resting exactly at the (logarithmic) centers of their cycles.

The scatter plot in Fig. 6.19 shows typical oscillations for digit 1 and focuses on the parametrical range of 0.05 to 57. This is obtained by successive computer simulations, which involve a tiny random factor as part of the scheme. Hence, the small fluctuations are noticed here but they are random in nature, errors, and lack of precision, not signifying anything systematic. Calculating directly the distributions is a much more difficult and involved task, besides being quite time-consuming. Such direct calculation should show an almost identical curve here, and it should be totally smooth.

Superimposing in one chart as shown in Fig. 6.20 excess leadership for digits 1, 2, 5, and 8, over and above their logarithmic centers, greatly helps in visualizing what is occurring here. The chart clearly demonstrates that it would be practically futile to compare parameters and search for those yielding (slightly) more conformity to the logarithmic. There exists no parameter here which can find them all simultaneously logarithmic. As we move anywhere on the parametrical  $p$ -axis, some digits get closer to their logarithmic values and some get farther away.

Magnitudes of oscillations are naturally higher for low digits and lower for high ones. The table in Fig. 6.21 shows the differences between the percent

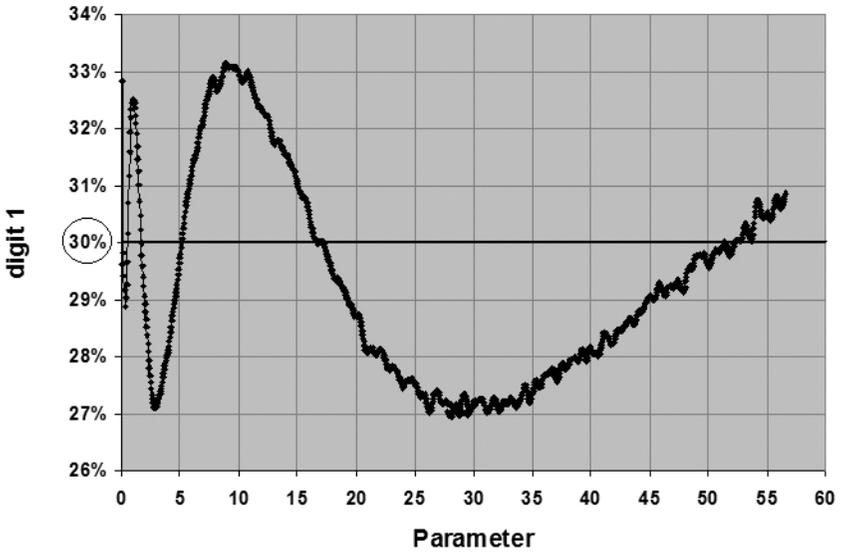


Figure 6.19 Digit 1 Leadership as a Function of Parameter — Exponential Distribution

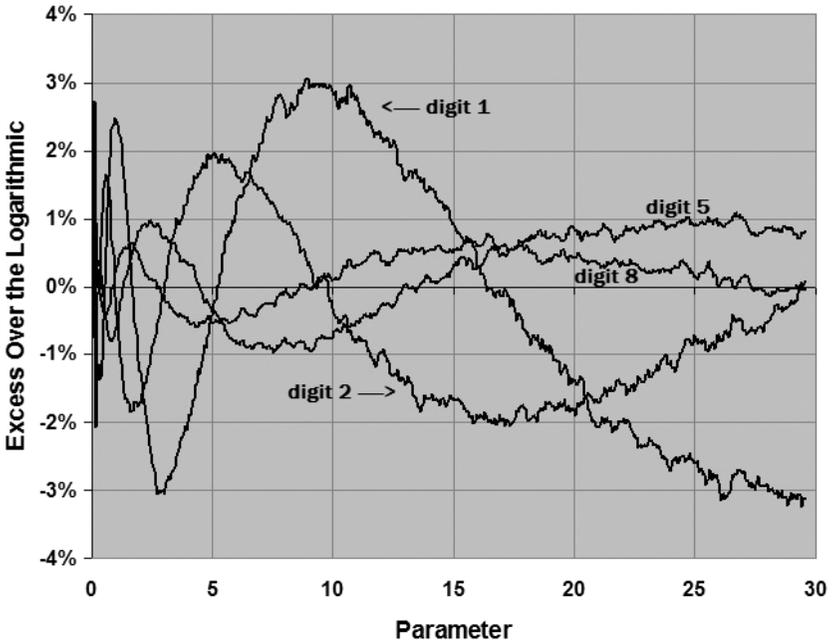


Figure 6.20 Excess Leadership for Digits 1, 2, 5, 8 over their Logarithmic Centers

Digit 1	Digit 2	Digit 3	Digit 4	Digit 5	Digit 6	Digit 7	Digit 8	Digit 9
6.2%	4.3%	3.3%	2.8%	2.2%	2.0%	1.7%	1.6%	1.5%

**Figure 6.21** Magnitude of Maximum Oscillation in the Exponential for Each Digit

leadership at the crests and the percent leadership at the troughs for all the digits (i.e. the range of variation for each digit).

A formal mathematical proof of the above results regarding the digital patterns of the exponential distribution was given by Christoph Leuenberger and Hans-Andreas Engel.

The chain **exponential(Uniform(0, M))** is simply a collection (aggregation) of diverse exponential cases where parameter  $p$  takes on various values within  $(0, M)$ . We have seen earlier the reasons why such a scheme should bring us very close to the logarithmic based on general arguments regarding the nature of chaining and the generic mechanizations at work here of fixed lower bounds and stretching/varying upper bounds. Figures 6.13 and 6.14 facilitate the visualization of the process involved in terms of the chain idea. Furthermore, in the case of this particular chain involving the exponential, one can also employ a totally different argument on top of the usual explanation, namely the argument regarding the beneficial role that the chain plays here in its function in smoothing out all the oscillations and deviations of all the nine digits from their respective logarithmic centers. The net effect of the act of chaining here is to average out distinct digital configurations and thus to cancel much of the deviations. This is a fertile ground for logarithmic convergence; the stage is always set in the exponential case for such near-perfect cancellations [before the introduction of chaining] since deviations are so strongly focused on  $\text{LOG}(1 + 1/d)$  for each digit, oscillating exactly around this value. Indeed, computer simulations show that overall digit distribution of such a chain of the exponential tied up to the Uniform is extremely close to the logarithmic far more so than any particular manifestation of an exponential with one particular (fixed) parameter  $p_0$  can ever aspire to be. These two separate arguments of why chaining here leads to the logarithmic are consistent and in harmony with each other, leading to the same conclusion. Interestingly, regardless of how carefully  $M$  is chosen in an effort to optimize logarithmic behavior, total logarithmic perfection can never be achieved here. This is due to the arbitrary nature of the selection of  $M$  which always slightly favors those digits whose cycles end at or near the value of  $M$ . For example, for  $M = 17$ , digit 1 benefits the most, since a single exponential with a fixed parameter of 17 is indeed logarithmic digit-1-wise.

Digit 2 on the other hand would have preferred values such as 3, 9, or 29, where it rests at its logarithmic proportion of 17.6%. Since M can only attain one single value, it cannot please everyone, and thus M is offered that distressing choice of befriending a particular digit while alienating all the other digits, and this is exactly why complete logarithmic behavior is not possible even here.

On the other hand, the chain **exponential(Lognormal(high shape))** is as perfectly logarithmic as can possibly be achieved for all practical purposes with an (almost perfectly logarithmic) Lognormal having, say, shape parameters larger than 1.5. This result of course is derived from the second conjecture of the chains of distributions. In (slight) contrast, the chain **exponential(k/x on (10, 100))** is perfectly logarithmic as it stands [formally with all the mathematical rigor and harshness one desires], since such  $k/x$  is exactly logarithmic.

## SAVILLE REGRESSION MEASURE REVISITED

---



---

Adrian Saville's novel approach in assessing the relationship between the first digits of a given data set and the Benford proportions is independent of the data size  $N$ , hence it can serve only as a comparison test at best, but definitely not as a compliance test. The algorithm is based on the classic simple linear regression model  $\text{Observed} = \mathbf{b} \cdot \text{Benford} + \mathbf{a}$ , where  $X$ , the independent variable, represents the theoretical Benford proportions and  $Y$ , the dependent variable, represents the actual observed proportions for the data set under consideration. Parameter  $\mathbf{b}$  is the slope and parameter  $\mathbf{a}$  is the intercept.

Let us apply Saville's method to the U.S. Census data on population of all cities and towns in 2009. First-leading-digits distribution here is:

**Data A:** {29.4%, 18.1%, 12.0%, 9.5%, 8.0%, 7.0%, 6.0%, 5.3%, 4.6%}.

The first digit proportions of the population data are designated  $Y$ , and it's regressed on the Benford proportions  $\text{LOG}(1 + 1/d)$  designated as  $X$ . The scatter plot alongside the resultant regression line minimizing sum of squared errors is shown in Fig. 6.22.

Closeness to Benford in this context necessitates that (I) the slope  $\mathbf{b}$  would be very close to 1, and that (II) the intercept  $\mathbf{a}$  would be very close to 0. As will be shown and discussed in detail towards the end of this chapter, each of these two requirements turned out to be equivalent to the other since slope and intercept have a one-to-one correspondence here, implying that it is enough to focus on just one measure. Recalling that the U.S. Census data on population centers adheres to Benford's Law very closely, we then expect to get a strong signal from Saville regression method that it actually does. Regression here yields (I) a slope of 0.9735, which is indeed quite near the ideal value of 1, and (II) an intercept of 0.0029 which is sufficiently near the ideal value of 0. The important thing to notice here is the close correspondence between the regression line (shown in black) and the  $45^\circ$  line  $Y = X$  (the longer gray line). The two lines almost overlap, signaling strong similarity to the logarithmic. Note that for this population data, SSD for the first order gave the extremely low value of 1.3, also signaling closeness to Benford.

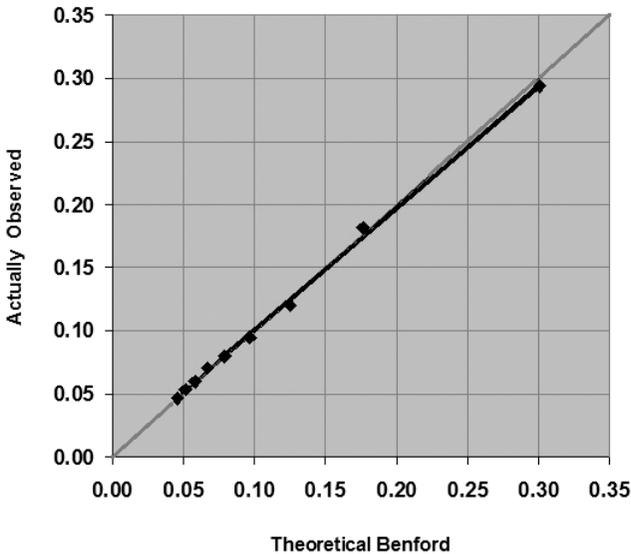


Figure 6.22 (Data A) Saville Plot for U.S. Population of Cities and Towns Data

Yet, closeness to Benford cannot be measured just in terms of slope and intercept alone. Another factor to be considered, and just as essential, is the variation of the actual points around their regression line. To illustrate this point we consider another set of hypothetical data.

**Data B:** {27.5%, 23.1%, 13.6%, 7.5%, 7.0%, 6.4%, 6.1%, 5.3%, 3.5%}.

Figure 6.23 depicts Saville regression plot where both lines actually coincide, since the slope of the (black) regression line is exactly 1, and the intercept is exactly 0. Yet, points scatter widely around their regression line in this data set. Surely Data B is farther from the Benford condition in comparison with Data A on U.S. populations, in spite of having that ideal slope of 1 and intercept of 0. Note that for Data B, SSD came out 45.2, signaling not-negligible deviation from the logarithmic.

Another example of Saville regression method is depicted in Fig. 6.24 for hypothetical data set that appears somewhat different than the logarithmic.

**Data C:** {24.6%, 14.6%, 11.3%, 9.7%, 9.0%, 9.1%, 7.9%, 6.3%, 7.5%}.

Digit distribution here is clearly less skewed in favor of the low digits compared with the Benford condition, and this fact can be clearly visualized by way of the flatter slope of the black regression line. Regression yields a slope of 0.674 and an

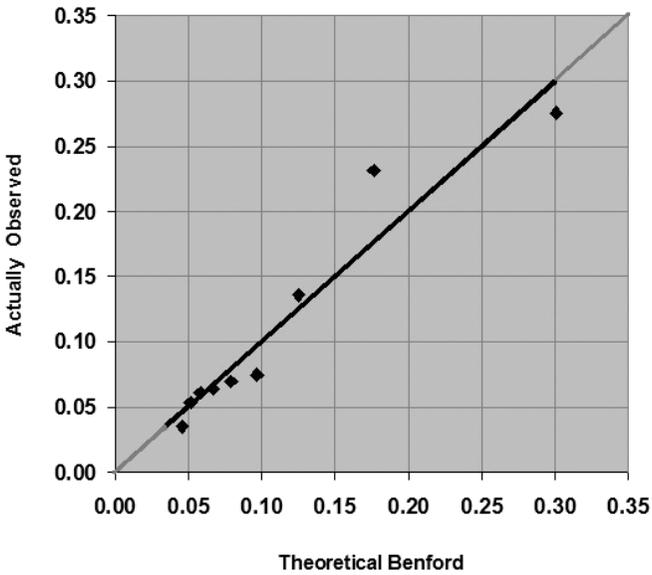


Figure 6.23 (Data B) Saville Plot — Ideal Slope 1 and Intercept 0 but with Deviations

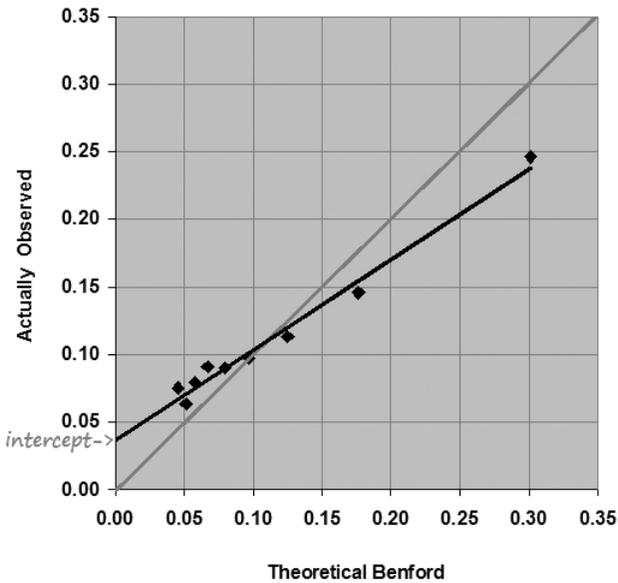
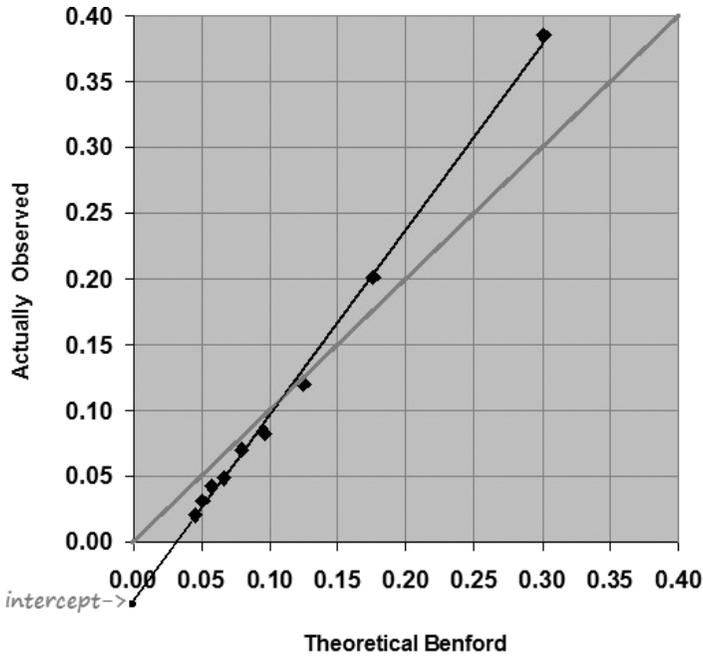


Figure 6.24 (Data C) Saville Plot — Flatter Slope and Positive Intercept



**Figure 6.25** (Data D) Saville Plot — Steeper Slope and Negative Intercept — Steady

intercept of 0.0362, indicating that the data set is not very close to the logarithmic. Also, SSD value of 62.1 for this data set is considered a bit too high.

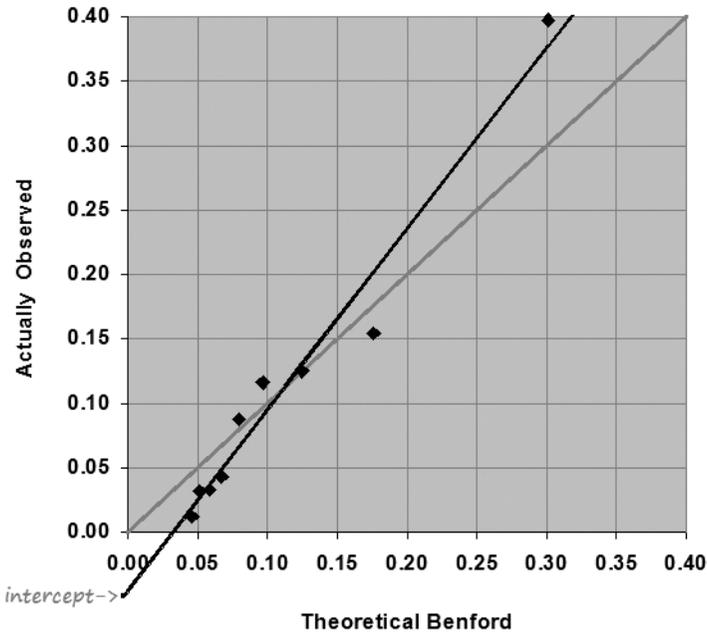
Another example of Saville regression method is depicted in Fig. 6.25 for hypothetical data set with highly skewed first-digit distribution, strongly favoring low digits.

**Data D:** {38.5%, 20.1%, 12.0%, 8.3%, 7.0%, 4.8%, 4.2%, 3.1%, 2.0%}.

Here digits 1 and 2 strongly dominate, even more so than in the Benford condition.

Regression yields a slope of 1.42 and an intercept of  $-0.0461$ , which indicates that the data set is not close to the logarithmic at all. For this data set SSD yields the value of 96.6 which also indicates not-negligible deviation from the Benford condition. One striking feature of this data set is the close and loyal adherence of almost all nine points around their black regression line, as if the data signals a strong determination and insistence on deviating from the Benford condition.

Figure 6.26 depicts Saville plot for another data set having almost the same slope and intercept as in Data D of Fig. 6.25, and with highly skewed first-digit



**Figure 6.26** (Data E) Saville Plot — Steeper Slope and Negative Intercept — Varied

configuration strongly favoring low digits, albeit with points scattered about indecisively and disloyally about their regression line, as if not showing enough determination in their behavior.

**Data E:** {39.7%, 15.4%, 12.5%, 11.6%, 8.8%, 4.3%, 3.3%, 3.2%, 1.2%}.

Regression yields a slope of 1.40 and an intercept of  $-0.0449$ , both of which are extremely close to the values of the previous example of Data D. For this data set SSD gave the value of 128.4, indicating a stronger deviation from the Benford condition, even more so than for data D, reflecting the extra variation around its regression line. The main difference between data sets D and E is how closely points adhere to their regression line.

Which data set should be considered farther from Benford and which should be considered closer, Data D or Data E? Do they both perhaps have an equal measure of deviation? On one hand, Data D is consistently adhering to that non-logarithmic regression line, having all its points doggedly and tenaciously marking their non-Benford path, while Data E is hesitant in doing so and thus demonstrates some flexibility in its anti-Benford stand. On the other hand, Data E shows more overall variability from the  $45^\circ$  Benford line itself, as evident by its higher 128.4

score of SSD, while Data D varies overall less around the 45° Benford line, as evident by its relatively lower 96.6 score of SSD. As a rule, it can be shown mathematically that the more the points vary and spread about their regression line, the more they deviate also from the 45° Benford line itself (as measured by SSD score, and other similarly defined measures).

Saville suggested sticking strictly to the slope and intercept output in such decisions, and in addition determining the measure of support the points lend their regression line in terms of the degree they adhere to it. Such a stand favors Data E, which blurs its anti-Benford message of slope and intercept being different than the ideal 1 and 0, respectively. So much so that he goes a step further here and advocates utilizing the probabilistic aspects of classic regression theory to decide on Benfordness of the data based on whether or not 1 and 0 are within the corresponding, say, 5% confidence intervals for the values of the slope and the intercept. In other words, Saville is suggesting applying the scheme not only as a comparison measure, but as a compliance test as well.

In essence, Saville's suggestion is not to focus at all on the relationship between the actual observed points and the gray 45° ideal Benford line, but rather on the similarity or dissimilarity between the black regression line  $Y = b \cdot X + a$  and the gray 45° ideal Benford line  $Y = X$ , as well as on the (tight or loose) relationship between the points and their regression line, to determine how much credence they give their resultant deviating slope.

Two extreme counter-examples to Saville's suggestion of using confidence intervals in the supposed context of compliance test should clearly illustrate the dilemma here.

Counter Example I: {30.4%, 17.8%, 12.6%, 9.7%, 7.9%, 6.6%, 5.6%, 5.0%, 4.4%}

Counter Example II: {21.7%, 36.8%, 9.6%, 14.5%, 1.0%, 1.0%, 3.4%, 6.5%, 5.5%}

Surely, the consensus here should give a decisive preference to Example I over Example II. Yet, even though Example I earned the near-perfect slope of 1.019, its 5% narrow confidence interval (considering slope alone with no regards to intercept) is calculated to be {1.012, 1.026} which does not contain the value 1.00, and thus the null hypothesis of Benfordness is rejected. The data is being punished for having its points follow very closely their slightly steeper regression line, as if insisting on a tiny bit deviation from Benfordness. Example II, on the other hand earns the perfect slope of 1.000 while showing more variation; its 5% wide confidence interval of {0.093, 1.907} easily contains 1.00, and thus readily passes the

test — its huge spread totally forgiven! In essence, Saville is disposed of excusing those huge deviations from Benford in Example II as being random in nature while viewing those tiny deviations in Example I as something more systematic!

Statistical criticism of such applications of the scheme is that standard errors and confidence intervals in classic regression theory are based on the assumption of normality and independence of the errors. Normality perhaps can be excused or explained away in our context, but not independence. To see why, imagine a single number within the data being changed from 150 to 897, thus increasing digit 8 proportion, while at the same time reducing digit 1 proportion. This illustrates the dependencies between the digits and the fact that no proportion can be changed without affecting the others. The fact that the sum of all digital proportions must add to 1 implies dependency between the proportions. Yet, the strongest criticism is voiced here at the very idea of giving confidence interval via a measure that does not incorporate  $N$  — the size of the sample/data. Since any supposed statistical setup of Saville's scheme for use as a proper compliance test would surely envision picking up a random sample (our data) from some (imaginary) very large and generic logarithmic population, this in turn implies that the size  $N$  greatly affects leading-digit distribution of the sample as measured by the slope, and so it must be incorporated into the statistic, when in fact  $N$  is nowhere to be found in Saville's scheme! Digital distribution of a truly small sample size, say, only 100 values randomly picked from some very large and perfectly logarithmic population of values, would typically yield result that varies wildly, and slope could easily become +5, 0, -4 and so forth. A sample size of just seven numbers implies that always at least two digits get the proportion 0! While an extremely large sample size would almost guarantee compliance, with slope closing in tightly around +1 in almost all samples and trials. Saville's scheme on the other hand always focuses on number 9, not on number  $N$ , as it addresses those resultant nine points of digital proportions on its scatter plot regardless of data/sample size  $N$ . It is certainly possible to conceive that a future rigorous statistical study incorporating  $N$  could give exact confidence intervals for the resultant slope of a well-defined statistical process in which a random sample (data) of size  $N$  is taken from a very large and generic logarithmic population, endowing resultant slope a particular distribution form. If mathematical statistics cannot come up with such a distribution form for the slope in the above process, then Monte Carlo simulation could shed some light on how slope is distributed. As an example, 2000 values are selected randomly from the set of 19,509 values of U.S. population centers data,

first digits calculated, and a singular value of Saville slope is then obtained. Such a process is repeated, say, 30,000 times, yielding 30,000 slope values which are then plotted as histogram in order to suggest the most appropriate distribution form enabling us to create (approximately) confidence intervals and cutoff points [but which would only be applicable for 19,509 population size and 2000 sample size, and assuming the digital configuration of U.S. population data which deviates slightly from the logarithmic].

The difficulties regarding the application of probabilistic confidence interval aside, another important issue to ponder here regards the researcher's proper focus. Should it be the slope, the intercept, or both? Saville chooses to view the null as double hypotheses, one about the slope being near 1, and the other about the intercept being near 0, and consequently he suggests applying Bonferroni joint confidence intervals. Yet, Saville's treatment of the slope and the intercept as two independent variables would be correct only for typical regression models where regression line is free to twist and turn and take on any slope and any intercept a priori, but not in the context of Benford's Law where resultant slope and intercept are restricted to proportion sets having total probability of 100%. Here, the slope and the intercept of the resultant regression line are always intrinsically constrained; they interlock. A given slope uniquely determines an intercept, and vice versa. A casual glance at Figs. 6.24 and 6.25 would convince anyone that (1) a slope flatter than 1 means intercept is positive and above the origin (for if intercept is at or below the origin then all nine points are dragged below the  $45^\circ Y = X$  line and their sum falls below 100%), and (2) a slope steeper than 1 means intercept is negative and below the origin (for if intercept is at or above the origin then all nine points are pulled above the  $45^\circ Y = X$  line and their sum rises over 100%). Let us prove this assertion by actually deriving an explicit algebraic expression for the intercept, assuming a given slope for the regression line. This is done based on a generic sketch of Saville plot shown in Fig. 6.27 where a flatter slope below 1 is assumed. The proof is to be generalized later for the other case where a steeper slope over 1 is considered. Let us recap: a slope flatter than 1 obviously cannot pass through the origin having 0 intercept, because that would mean that each digital proportion is less than its associated Benford one and thus total proportion is less than 100%. Rather, a flatter slope implies that the line must pass above the origin having a positive intercept. In the same vein, a slope steeper than 1 cannot pass through the origin having 0 intercept, because that would mean that each digital proportion is more than its associated

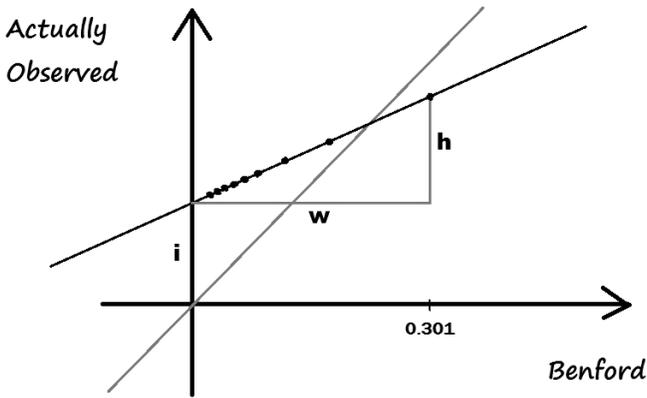


Figure 6.27 A Generic Sketch of Saville Plot — Relating the Slope and the Intercept

Benford one and thus total proportion is more than 100%. Rather, a steeper slope implies that the line must pass below the origin having a negative intercept. One basic linear regression fact to be utilized here states that the sum (or the average) of all Y coordinates on the regression line itself is equal to the sum (or the average) of all Y coordinates of the actually observed points. Here, the sum of all observed Y values is 1, namely 100%, and so should be the sum of all Y values/coordinates on the regression line itself.

Note that the height *i* on the sketch signifies the intercept. By definition, slope =  $h/w$ . For digit 1,  $w = 0.301$ , and  $h_1 = \text{slope} * 0.301$ . For digit 2,  $w = 0.176$ , and  $h_2 = \text{slope} * 0.176$ . Since each  $Y_d$  coordinate is  $(i + h_d)$ , summing up all nine Y coordinates of the points on the regression line entails summing up  $(i + h_d)$  for all nine cases. We get:

$$\begin{aligned}
 1 &= [i + h_1] + [i + h_2] + \dots + [i + h_9] \\
 1 &= [i + \text{slope} * 0.301] + [i + \text{slope} * 0.176] + \dots + [i + \text{slope} * 0.046] \\
 1 &= [9 * i] + [\text{slope} * (0.301 + 0.176 + \dots + 0.046)] \\
 1 &= [9 * i] + [\text{slope} * (1)] \\
 1 &= [9 * \text{intercept}] + \text{slope}
 \end{aligned}$$

$$\text{intercept} = (1 - \text{slope}) / 9$$

where both the slope and the intercept here refer to the regression line. If we consider the fact that 1 and 0 are the slope and the intercept, respectively, for

the ideal Benford line  $Y = X$ , then in order to express an alternative interpretation, this can be written more elegantly as:

$$\begin{aligned}(0 - \text{intercept}) &= -(1 - \text{slope})/9 \\ (\text{Intercept}_{\text{Benford}} - \text{Intercept}_{\text{Observed}}) &= (-1/9) * (\text{Slope}_{\text{Benford}} - \text{Slope}_{\text{Observed}}) \\ 9 * (\text{Intercept}_{\text{Benford}} - \text{Intercept}_{\text{Observed}}) &= -(\text{Slope}_{\text{Benford}} - \text{Slope}_{\text{Observed}})\end{aligned}$$

In words: “Regarding the resultant Saville regression line for any observed data: for each unit of positive intercept difference from the Benford line there is a ninefold difference in the slope having an opposite sign”. (That ninefold value emanates from having nine digits).

For the case with a slope steeper than 1, the exact same result is gotten. The intercept value  $i$  in the sketch is now necessarily negative. On each of the nine points on the regression line,  $w$  is the same as in the previous sketch, but the total height relating to slope definition is the  $Y$  coordinates plus the absolute value of  $i$ , hence

$$\begin{aligned}\text{slope} &= [(Y \text{ coordinate}) - i]/[w], \text{ and therefore } [Y \text{ coordinate}] = [w * \text{slope} + i]. \\ 1 &= [\text{all } Y \text{ coordinates}] \\ 1 &= [0.301 * \text{slope} + i] + [0.176 * \text{slope} + i] + \dots + [0.046 * \text{slope} + i]\end{aligned}$$

Hence  $1 = (1) * \text{slope} + 9 * i$ , and solving for the intercept  $i$  we get the same expression as in the other scenario, namely  $\text{intercept} = (1 - \text{slope})/9$ , but now since the quantity  $(1 - \text{slope})$  is negative, so is the intercept! In the other scenario, for a slope flatter than 1, the quantity  $(1 - \text{slope})$  is positive, hence the intercept is positive, namely hanging above the origin.

The table in Fig. 6.28 depicts a variety of different digital configurations, from one having extreme digital skewness in favor of high digits, all the way to another extreme case favoring low digits, and a few cases in between. Saville regression values for slope and intercept are calculated in each case. The above succinct and general result relating slope and intercept is used in calculating the intercept once again via  $(1 - \text{slope})/9$  as the last row at the bottom. This last row perfectly agrees with that of the (regression-calculated) intercept in the other row just above it, confirming our mathematical reasoning above.

Which measure should be considered preferable or more appealing here, the slope or the intercept? Each variable is simply a different manifestation of that slope-intercept duality measuring degree of steepness, yet the slope seems much more natural and quite familiar given its prominent role in calculus expressing the

DIGIT	Extreme skewness (high)	Almost extreme skewness (high)	Skewed in favor of high digits	Digital Equality	Less skewed than Benford	Benford	Skewer than Benford	Almost extreme skewness (low)	Extreme skewness (low)
1	0.0	0.0	4.7	11.1	24.7	30.1	43.1	60.0	100.0
2	0.0	0.0	5.1	11.1	14.7	17.6	22.3	40.0	0.0
3	0.0	0.0	6.5	11.1	12.3	12.5	11.2	0.0	0.0
4	0.0	0.0	7.6	11.1	10.5	9.7	7.3	0.0	0.0
5	0.0	0.0	9.8	11.1	9.9	7.9	5.1	0.0	0.0
6	0.0	0.0	11.9	11.1	8.8	6.7	4.1	0.0	0.0
7	0.0	20.0	12.0	11.1	7.6	5.8	3.3	0.0	0.0
8	0.0	30.0	17.7	11.1	6.1	5.1	2.4	0.0	0.0
9	100.0	50.0	24.7	11.1	5.3	4.6	1.2	0.0	0.0
Slope	-1.20	-1.13	-0.54	0.00	0.71	1.00	1.65	2.58	3.49
Intercept	0.245	0.236	0.171	0.111	0.032	0.000	-0.072	-0.175	-0.277
(1-Slope)/9	0.245	0.236	0.171	0.111	0.033	0.000	-0.072	-0.175	-0.277

Figure 6.28 Slope and Intercept for Various Cases, Plus  $(1 - \text{slope})/9$  Confirmation

same concept. Saville setup cannot provide probabilistic measure of compliance, yet it has led to an elegant and concise measure of steepness, or rather to a measure of ‘skewness over and above the Benford condition’. That concept can now be succinctly defined and measured via the single variable of the slope which designates data as ‘more skewed than Benford’ whenever it is larger than 1, and ‘less skewed than Benford’ whenever it is less than 1. Such a concise and clear measure turned out to be immensely useful in measuring digital development patterns for all random data types, especially so in light of the fact that here there is no need to decide which digits are to be termed ‘low’ and which are to be termed ‘high’. The statement “Digital configuration is skewed in favor of low digits even more so than the Benford condition” formally necessitates breaking down our nine digits into two competing groups in a reasonable yet arbitrary manner. For example, low digits could be designated as  $\{1, 2\}$ , and high ones as  $\{3, 4, 5, 6, 7, 8, 9\}$ . Perhaps a ‘better’ partitioning could be found in pitting  $\{1, 2, 3, 4\}$  against  $\{5, 6, 7, 8, 9\}$ . How do we go about it, and what is exactly meant by a ‘better partitioning’ mathematically? Saville method enables us to avoid this dilemma altogether. Digital development implies that Saville’s slope on those mini sub-intervals between adjacent IPOT points experiences the transition from being either negative or near

zero on the left, to about 1 around the center, and then rising above 1 on the far right part of the range. This transition of the slope can then be quantified in the approximate to describe the degree of development occurring throughout the entire range, and a single number signifying a measure of development can then be found. It is hoped that this single number is roughly constant across all types of random data, or that at least its variance is very low. In other words, that there exists approximately uniform measure of development across all random data (logarithmic or otherwise) so that a universal law of development could be formally stated via that singular numerical value. At a minimum it is hoped that such a uniform measure of development exists for all logarithmic random data sets.

In summarizing, Saville regression method not only cannot be applied as a compliance test, but it also cannot be applied as a comparison test. Yet, it can be applied as an elegant and efficient measure of development. Knowing that Saville's slope is exactly 1 and that the intercept is exactly 0 as seen earlier in the Counter Example II of {21.7%, 36.8%, 9.6%, 14.5%, 1.0%, 1.0%, 3.4%, 6.5%, 5.5%} does not imply that digits are really anywhere near the logarithmic, as evident from a cursory look at this distribution. The same difficulty is easily visualized in Fig. 6.23 for Data B of {27.5%, 23.1%, 13.6%, 7.5%, 7.0%, 6.4%, 6.1%, 5.3%, 3.5%}, which comes with a perfect slope of 1 and yet deviates somewhat from the logarithmic. It is only when slope is not anywhere near the ideal Benford value of 1 (say, slope is 3 or  $-2$ ) that one may deduce that the given data set is not logarithmic at all.

A cursory look at Saville scatter plot suggests that digits 1 and 2 play a very big role overall in the algorithm. The position on the probability axis of 0.301 and 0.176 for digits 1 and 2 respectively, compared with the congregation of high digits near the origin necessarily implies that 1 and 2 dominate resultant slope/intercept, a fact that is obvious to anyone with some experience in applying simple linear regression. One has to keep in mind that it is precisely the deviation in the proportions of digits 1 and 2 that typifies non-conformity, as higher digits must 'passively' respond by deviating in the opposite direction. Since the law has granted digits 1 and 2 the combined large proportion of 47.7%, which is roughly a half, naturally they play a larger role in the overall scheme of things than do the other digits. In other words, digits 1 and 2 are granted such prominence in swaying results here simply because they were intentionally and deliberately put way ahead of the others to the far right on Saville probability axis by none other than Frank Benford himself!

All in all, the setup of Saville's regression scatter plot is very attractive visually. The scatter plot really tells the whole story of data-to-Benford-comparison in a very concise manner. The digital analyst should focus on the following two aspects of the scatter plot: (I) the correspondence or difference between the regression line and the  $45^\circ Y = X$  theoretical Benford line, representing the line that would have been gotten had data been perfectly logarithmic, and (II) the degree individual points deviate from their regression line.

An additional suggestion is made regarding an extension of Saville method for the second-order leading digits. There though, the range of variability is narrowly focused on the very short interval  $(0.085, 0.120)$ , making the slope all too sensitive to even some minute deviation from Benford. Therefore it might be suggested to transform all second-order proportions by an adjustment left translation (subtraction) of 0.085, so that a proportion of 0.085 is entered as 0, and for the slope to be defined in the usual way. Another possible extension may be the application to the first-two-digits distribution, where the range varies nicely, widely, and gradually from 0.0044 all the way to 0.0414, and where robustness of the slope is more likely to be found due to the great abundance of 90 points. A further extension of Saville method is suggested in its very general application in mathematical statistics — measuring deviations for all discrete probabilistic distributions, such as the binomial and Poisson distributions.

## THE SCALE INVARIANCE PRINCIPLE AND AGD INTERPRETATION

---

---

Derivation from the scale invariance principle has been criticized for making unfounded assumptions. Nonetheless, let us justify the scale invariance principle by basing the argument on some very general consideration about the nature of typical usage of numbers in the context of the Aggregate Global Data Interpretation of the law. Hill's distribution of all distributions model and its related AGD Interpretation of the law serve here as the backdrop of the following discussion, where the law is stated in relation to the totality of all our recorded data.

Scales (units) are arbitrary measuring physical rods against which we compare other entities. We define here the term **'measured category'** as some very specific physical entity of a particular category that needs measuring, such as heights of people, heights of horses, heights of buildings, lengths of rivers, distances between global cities, ages of people, or lifespan of bacteria (these examples constitute seven different measured categories in all). Scales are used in quoting measured categories. Let us assume for a moment that society has standardized all measurements and does not tolerate using different scales for different measured categories of the same dimension, and that only a single universal scale is used to measure all measured categories regarding any one particular dimension (for example, for the time dimension, only the second is used while hours, minutes, days, years, nanoseconds, and months are totally excluded).

On the face of it, the scale invariance principle appears totally unfounded. For example, people's height measured in feet yields mostly 5 and 6 as the leading digits while on the meter scale digit 1 takes a strong lead. Yet, the rationale for the principle rests on the assumption that our measured categories are not just numerous, but enormous. It also rests on the notion that changes in scale will have such varying and independent effects on measured categories in terms of leading digits that it will all sum up to nothing. In other words, even though a change in scale would revolutionize digital leadership for almost each and every measured category (individually), yet the changes for almost all measured categories do take

totally different turns, canceling out each other's effects and leaving the net results on digital leadership unaffected. Had we used only four measured categories for, say, the length dimension, then a change in scale would certainly not thoroughly shuffle/transform all numbers in such a way as to leave digital leadership unaltered. But we have the use of many more measured categories of length than merely four. Intuitively, any single change of scale that would result in digit 7, for example, taking a stronger lead all of a sudden in a certain measured category would be offset somewhere else (another measured category) where digit 7 would be all at once deprived of some existing leadership, balancing things out. All this is relevant and correct if Benford's Law is applied and interpreted regarding the totality of everyday data, not to any subsets thereof or merely one measured category. As for specific example, consider two measured categories: heights of people in feet where most quotes are assumed to be on the interval (4.5 ft, 6.7 ft), and heights of buildings in a certain region where most quotes are on the interval (65 ft, 193 ft). If standard scale is changed to meters using the conversion formula (FEET measurements)\*0.30919 = (METER measurements), this would give roughly the new intervals (1.39 m, 2.07 m) and (20.1 m, 59.7 m). Clearly digit 1 gained a lot of leadership with people's height under this scale transformation, but lost badly with the buildings. This trade-off is due to measured categories being something real and physical, and that the two intervals in the example above relate to each other in a fixed manner independently of the scale chosen. The fact that digit distribution of the totality of all our measured categories remains unaffected under a unified scale change is fairly intuitive, yet a rigorous mathematical proof of this grand trade-off is sorely lacking.

It is quite intuitive that under the assumption of uniform scale for all measured categories, any possible 'digital leadership bias' about the choice of scale during earlier epochs when it was decided upon would be totally ineffective due to the enormous size of the number of measured categories now in existence. In other words, under the assumption of uniqueness of scale, society could never find one magical scale that would give leadership preferences to some digits at the expense of others.

We could relax the above assumption of uniform scale for all measured categories and even go to the other extreme by imagining a society that is extremely fond of multiple units, and in addition tends to assign units in haphazard and random fashion without any particular calibration with regard to leading digits or any other sensible manner, a society that liberally invents different scales for each and

every measured category. Given the assumption of the enormity of the number of measured categories, even for such an idiosyncratic and inefficient society the invariance principle still holds and the reshuffling of all their units in one single epochal instant (either uniformly by a fixed factor for all measured categories, or by different factors assigned randomly for each measured category) would not result in any change in overall leading-digits distribution. Such a society also passes the scale invariance test.

But let us consider a third case, of another society, one that is also very fond of multiple units and liberally invents different scales for each and every measured category yet doesn't assign units in haphazard and random fashion. Rather, it tends to carefully calibrate things in such a way so as to quote numbers with mostly digit 1 and 2 leading. For such a society, the scale invariance principle has no validity because the reshuffling of all the scales simultaneously would drastically alter leading-digits distribution (in favor of the logarithmic if done randomly and differently for all measured categories).

And what about our society during this particular epoch? Surely our society does limit diversity of scales a great deal. We don't liberally invent different scales for each and every measured category, yet we do not completely restrict ourselves either. It is also true that we do not usually assign units completely haphazardly and randomly when inventing new ones. There are numerous examples of convenient yet peculiar and drastic adjustments of scale, such as astronomical units of length being light years, Avogadro's number in chemistry, and many others. Typical examples are the scales for some measured categories regarding weights: tons are used at international freight companies, milligrams in biological laboratories, and kilograms at butcher shops, with the overall result that many quotes of weight are with low-digit-led numbers (1.34 ton of crude oil delivered, 3.5 milligram of  $C_6H_3O_5$  solution, 1.5 kilogram beefsteak). It might even be argued that people unknowingly and subconsciously like dealing with low-digit-led numbers, especially digits 1 and 2, and that they at times invent scales that promote such outcome. In any case, even if that is true in some instances, the fact that we severely limit multiple usages of scales compared with our huge number of measured categories implies that no matter how we calibrate scale for this or that particular measured category, inevitably numerous other measured categories (all united in paying homage to one single scale) will have the final say in overall leading-digits distribution and will swamp them by their sheer size. Hence our society (where Roger Pinkham grew up) easily passes the scale invariance principle.

Clearly, a change in scale for one particular dimension implies the transformation of numerous measured categories by different multiplicative factors. For example, a change in the length dimension that reduces standard unit to half its original size would require the factor of 2 for simple lengths, 4 for areas, 8 for volume,  $1/2$  for kilograms per length,  $1/8$  for minutes per volume, and so forth. Similar changes in the units of mass and time dimensions would result in different factors. Nonetheless, a simplistic version of the scale invariance principle can also be invoked just the same by stating that one single unified transformation of all measured categories (of the totality of our data) by any single multiplicative factor would not alter existing logarithmic distribution whatsoever. As a corollary, transforming logarithmically well-behaved and large enough pieces of data by multiplying them by any single factor would barely nudge their digit distributions. Readers are encouraged to attempt to perform this rather striking demonstration of the scale invariance principle.

## CASE STUDY XI: LARGE SAMPLE FROM A VARIETY OF DATA SOURCES

---

A large sample of real-life mixed data was collected as a rudimentary test for some of the ideas presented in this book. Results strongly confirmed the existence of a digital signature within a signature for Hill's scheme, namely the very definite and peculiar way digits develop their logarithmic property along their entire range of data. Also, the conjecture that the totality of everyday data [and, by extension, Hill's model] mimics the Lognormal distribution was somewhat confirmed, or at least not completely rejected.

The data was collected in the spirit of Ted Hill's distribution of all distributions model. Effort was put into obtaining only a small number of values from each particular source or topic while varying the topics as much as possible so as not to focus too much on any one data set. In total 34,269 positive values were obtained from 70 different sources or topics. Since some topics were related or similar in nature, this last value of 70 can be realistically approximated as 50 or 60. Year numbers such as 2001, 1976, and so forth were omitted deliberately since their frequency seemed exaggerated. Numbers representing codes, indexes, and pages, were also omitted. Surprisingly, negative numbers were few and far between, representing less than 0.5% of overall original/raw data collected, and were also omitted.

The sources of the collected data were from the following six websites:

*<http://www.ers.usda.gov/>*

U.S. Department of Agriculture, Economic Research Service.

Topics included: Animal Products, Crops, Diet, Health, & Safety, Farm Economy, Farm Practices & Management, Food & Nutrition Assistance, Food Sector, Natural Resources & Environment, Rural Economy, Trade & International Markets.

<http://www.census.gov/>

U.S. Census Bureau.

Mostly data about Population Statistics, as well as Employment, Health, and Business Dynamics Statistics.

<http://statisticsworldwide.com>

Worldwide Business data, Governmental data, and some General International Statistics.

<http://www.scotland.gov.uk/Disclaimers>

The Scottish Government Data Center regarding a large variety of issues.

<http://www.ncaa.org/>

National Collegiate Athletic Association, relating to Sport Statistics.

<http://data.worldbank.org/topic/health>

World Bank information regarding Health, Population, Nutrition, Life Expectancy, and General Economic Statistics.

First leading digits of the aggregate data set of 34,269 values came out very close the logarithmic, with low SSD value of 4.8, indicating that the data set is nearly logarithmic.

The first-order digit distribution is:

Mixed Numbers, Varied Source — {28.8, 16.4, 12.4, 9.8, 8.3, 7.3, 6.1, 5.7, 5.3}

Benford's Law First-Order Digits — {30.1, 17.6, 12.5, 9.7, 7.9, 6.7, 5.8, 5.1, 4.6}

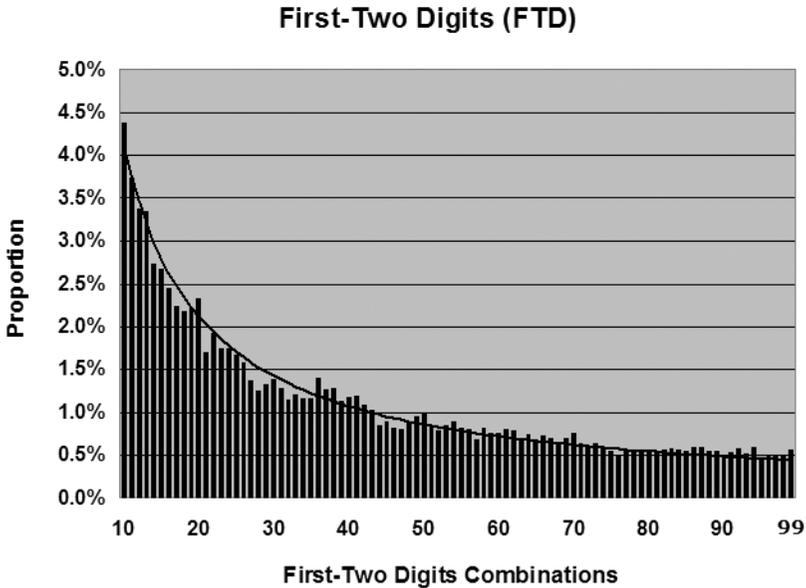
For the calculations of the second-order digits, 2,258 numbers without a second digit were omitted, leaving only 32,011 numbers in the calculations. The numbers omitted typically were such as 3, 8, 9, 0.1, 0.6, 0.07, 0.01, and so forth, having no meaningful second digit. The extremely low SSD value of 0.49 gotten here is rarely found in real random data, even in highly logarithmic ones.

The second-order digit distribution is:

Mixed Numbers — {12.4, 11.2, 10.8, 10.6, 9.9, 9.4, 9.4, 8.8, 8.7, 8.7}

BL 2nd Order Digits — {12.0, 11.4, 10.9, 10.4, 10.0, 9.7, 9.3, 9.0, 8.8, 8.5}

Examining first-two-digits (FTD) chart for those 32,011 numbers also confirms the strong logarithmic behavior of this data set, as shown in Fig. 6.29.



**Figure 6.29** First-Two Digits of Data Gathered From a Large Variety of Sources

Figure 6.30 depicts the histogram of related (decimal) log of all the data [34,269 values]. It suggests that the totality of our everyday data could perhaps be approximated by a Lognormal-like curve as was conjectured earlier. It is instructive to notice the relatively huge range on the log-axis here (about 10 units!) compared with the modest 2.5 or 3.0 length requirement of related log conjecture. It is extremely rare for any single-issue real-life random data set to come with such exceptionally high order of magnitude! Surely, such large order of magnitude is the result of appending and mixing numerous unrelated data sets of varying (small) orders of magnitudes [as is suggested in Fig. 4.46].

Examining digital development along the relevant sub-intervals standing between adjacent IPOT points [for the entire data set of 34,269 values] reveals a clear pattern as seen in Figs. 6.31 and 6.32. Around the left region of the data on  $[0.001, 1)$  where concentration is less than 5% of overall data, digital equality is approximately observed. On  $[1, 10)$  the data (12.6% portion) is mildly skewed in favor of low digits but not as skewed as the logarithmic. Around the center on  $[10, 1,000,000)$  the data is approximately logarithmic, and that is where most of the data occurs (75% of overall data). Farther to the right on  $[10^6, 10^9)$ , with only approximately 7% of overall data, more extreme digital inequality occurs. The last

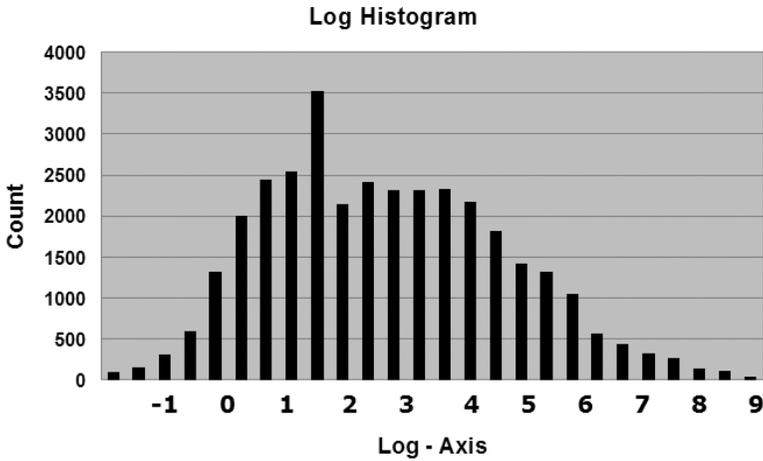


Figure 6.30 Histogram of Log of Data Gathered From a Large Variety of Sources

Left Point	0.001	0.01	0.1	1	10	100
Right Point	0.01	0.1	1	10	100	1,000
	=====	=====	=====	=====	=====	=====
Digit 1	9.7	10.6	17.8	21.8	27.6	27.7
Digit 2	22.6	16.7	12.4	15.1	13.4	16.1
Digit 3	9.7	12.2	10.2	14.4	12.2	12.4
Digit 4	16.1	10.2	9.0	11.8	9.6	9.1
Digit 5	3.2	8.6	7.9	9.1	8.8	8.9
Digit 6	16.1	7.3	9.5	9.0	9.2	6.9
Digit 7	12.9	9.4	9.4	6.3	7.7	6.3
Digit 8	3.2	11.4	11.4	6.5	5.6	6.9
Digit 9	6.5	13.5	12.6	5.9	6.1	5.7
% of Data	0.1%	0.7%	3.9%	12.6%	20.6%	16.4%
ES12	-15%	-20%	-18%	-11%	-7%	-4%

Figure 6.31 Digital Development — From Digital Equality to the Logarithmic

line at the bottom provides ES12 (Excess Sum digits 1 and 2) measure of skewness over and above the logarithmic condition. Variable ES12 shows a clear pattern of rising from around -15 to around +9. [Note: By convention, all left edges are open and all right edges are closed, as in [L, R), so as not to overlap any points. The tables in Figs. 6.31 and 6.32 were constructed in accordance with this convention].

Left Point	1,000	10,000	100,000	$10^6$	$10^7$	$10^8$
Right Point	10,000	100,000	1,000,000	$10^7$	$10^8$	$10^9$
	=====	=====	=====	=====	=====	=====
Digit 1	28.3	34.6	31.3	44.3	35.9	40.5
Digit 2	18.6	18.5	19.6	15.5	20.7	16.0
Digit 3	12.7	12.1	11.8	11.3	10.0	11.8
Digit 4	9.2	10.9	8.5	8.0	11.0	9.5
Digit 5	8.1	6.8	9.0	6.1	6.7	6.1
Digit 6	6.7	5.2	6.4	5.5	5.6	8.4
Digit 7	5.6	4.4	5.7	3.2	3.7	4.6
Digit 8	5.6	4.1	4.4	3.7	2.4	2.7
Digit 9	5.2	3.4	3.4	2.4	4.0	0.4
% of Data	15.0%	13.9%	9.2%	4.4%	2.0%	0.8%
ES12	-1%	5%	3%	12%	9%	9%

Figure 6.32 Digital Development — From the Logarithmic to More Severe Inequality

Also of note here is that both 0 and 1 serve as strong anchors for the data. In other words, the range immediately to the right of the 0 origin and 1 is where data congregates most, thinning out on the x-axis onwards, with a one-sided tail to the right falling off strongly from high concentration areas near 0 and 1. Here, there isn't any temporary initial rise in the density akin to the Lognormal with high shape parameter value. In this sense the data gathered here shows some similarity to the exponential distribution in a sense by consistently falling off from the very beginning around the origin.

Interestingly, random samples of only 1000 values from the total population space of 34,269 values yield better agreement with the logarithmic. This is so since such a sample better complies with Hill's model assumptions as it mixes values from a variety of sources better (i.e. fewer values per source), and this is so in spite of its limitation in having a much smaller data size of 1000 as compared with 34,269 of the population's size! In one typical example, 1000 picks from the well-shuffled 34,269 set gave excellent agreement with the logarithmic, where first digits came out {29.9, 17.0, 12.7, 9.1, 8.2, 7.7, 5.4, 5.1, 4.9} with the very low SSD value of 2.2 [lower than the 4.8 SSD value of the entire data set].

## DIRECT EXPRESSION OF FIRST DIGIT FOR ANY NUMBER — COMPUTER USE

---

A useful expression yielding the first digit of a positive number  $X$  is given by:

$$\text{1st Digit of } X = \text{INT}(X/10^{\text{INT}(\text{LOG}_{10} X)})$$

The INT function refers to the integer on the  $(-\infty, +\infty)$  x-axis immediately to the left of the number  $R$  in question. If  $R$  is exactly an integer then  $\text{INT}(R)$  is that same integer. If  $R$  is positive then  $\text{INT}(R)$  is just the whole part excluding the fractional part. If  $0 \leq R < 1$  then  $\text{INT}(R)$  is 0. If  $R$  is negative then  $\text{INT}(R)$  is the largest negative integer less than or equal to  $R$ . For example,  $\text{INT}(8.3)$  is 8,  $\text{INT}(4)$  is 4,  $\text{INT}(0.6)$  is 0, and  $\text{INT}(-3.7)$  is  $-4$ .

$\text{INT}(\text{LOG}_{10} X)$  is called the characteristic, namely the integral part of the log if  $X \geq 1$ . Since any number  $X$  can be broken down into  $X = 10^{\text{Characteristic}} 10^{\text{Mantissa}}$ , hence  $X/10^{\text{INT}(\text{LOG}_{10} X)} = 10^{\text{Characteristic}} 10^{\text{Mantissa}} / 10^{\text{Characteristic}} = 10^{\text{Mantissa}} = \text{Significand}$ .

Therefore the expression  $\text{INT}(X/10^{\text{INT}(\text{LOG}_{10} X)})$  for the first digit of  $X$  is actually also  $\text{INT}(\text{Significand})$ . And since  $\text{significand} \in [1, 10)$  and it is that first part of the scientific notation with the decimal location totally ignored,  $\text{INT}(\text{Significand})$  yields the first leading digit. This expression is immensely useful for computer implementations in Benford's Law. All computer programming languages contain an INT-type function. In MS-Excel, one of the best software choices for use in Benford's Law applications, the function  $= \text{INT}(\text{argument})$  is built into the system.

## ARTIFICIALLY CREATING NEARLY PERFECT LOGARITHMIC DATA

---

In order to create data that is nearly perfectly logarithmic via computer simulations or calculations, four methods shall be suggested here. It must be acknowledged though that  $\text{LOG}(1 + 1/d)$  is an irrational number and thus no finite data set can ever aspire to observe Benford's Law exactly. For example, for any data set constructed,  $[\text{numbers led by digit } 1]/[\text{total numbers created}]$  is a rational number, and can never be equal to  $\log(2)$  no matter how large the data set is made.

The four methods are written symbolically in the order they are explained below:

$e^{\text{Normal}(\text{any mean, s.d.} > 1)}$

$10^{\text{Uniform}(0, 1)}$

$10^{\{0.001, 0.002, \dots, 0.998, 0.999\}}$

$\{\text{IPOT}, \text{IPOT} * F, \text{IPOT} * F^2, \text{IPOT} * F^3, \dots, \text{IPOT} * F^N \approx \text{Much Higher IPOT}\}$

(1) *Simulate values from the Normal distribution with s.d. > 1, then take 2.718281 (namely e) to the power of those realized numbers.*

The result is a Lognormal distribution with a sufficiently large shape parameter and therefore nearly logarithmic. Using our decimal base 10 instead of e is even more effective! [In accordance with what was stated in the last paragraph of Chapter 64].

(2) *Simulate  $k/x$  over (1, 10) by way of taking 10 to the power of simulated values from the Uniform on (0, 1).*

By Proposition I, the log of  $k/x$  is Uniform, and by Corollary I the process is logarithmic if Uniform is spread over an integral length. A much more direct and simplistic approach here is to view this process as creating log values on (0, 1) via the Uniform, which in turn means the creation of uniform mantissa over (0, 1)!

(3) *Artificially create evenly spaced values (log-wise) of  $k/x$  over (1, 10) by way of taking 10 to the power of  $\{0.001, 0.002, 0.003, \dots, 0.997, 0.998, 0.999\}$ .*

If better accuracy and larger data set is desired, this can be further refined by taking 10 to the power of  $\{0.00001, 0.00002, 0.00003, \dots, 0.99997, 0.99998, 0.99999\}$  for example.

(4) *Create a non-rebellious exponential growth series.*

Ideally the series should start from an IPOT value and ends approximately near another much higher IPOT value, so as to satisfy the two requirements of sufficient length and an integral exponent difference.

To avoid rebelliousness, simply avoid all  $F$  such that  $\text{LOG}_{10}(F) = L/T$  with both  $L$  &  $T$  being integers. Moreover, avoid having  $\text{LOG}_{10}(F)$  being even anywhere near a rational number. Ensuring that  $\text{LOG}_{10}(F)$  is less than (approximately) 0.01 would guarantee that even if by mistake it was constructed near a rational number, deviation from the logarithmic is still minimal as mentioned in Chapter 98 and demonstrated in Fig. 6.3.

The first two suggestions (1) & (2) are random (simulations), and the last two (3) & (4) are deterministic (calculations). Yet it should be noted that only the first one of the Lognormal is of the random flavor in Benford's Law as it is spread over many IPOT intervals and shows digital development, while (1), (2), & (3) are of the deterministic flavor having flat/uniform related log density, and a consistent logarithmic digital behavior throughout. Method (3) should produce the closest result, the nearest one can get to the logarithmic, provided that enough refinement in incremental width is applied, 0.0000001 for example. The first two methods (1) & (2) always involve some tiny random error in simulation. The challenge in the last method (4) is to carefully calibrate things so that the starting and ending elements in the series are both almost exactly near IPOT values.

## **Section 7**

### **THE LAW OF RELATIVE QUANTITIES**

**This page intentionally left blank**

## THE RELATING CONCEPTS OF DIGITS, NUMBERS, AND QUANTITIES

---

In this last section, Benford's Law, a supposedly digital phenomenon, is found and demonstrated to be a consequence of something much larger, the result of a much more general and universal law governing quantities. The focus then naturally shifts, from the strong attention given previously to the relative occurrences of digits, to the investigation of the relative occurrences of quantities.

If the law is also (or mostly) about quantities, then it seems quite curious that we must resort to digits in order to express the phenomenon. On the face of it, there is no obvious connection between digits and quantities. Aren't digits merely invented symbols enabling us to express numbers, and by extension to convey quantities? Couldn't the law be stated directly regarding quantities without any digital involvement? Is it merely a mathematical coincidence perhaps that in order to express this quantitative phenomenon we must resort to digital distributions and that there is no other way to accomplish this? In other words, could it be that by pure chance we are blessed with a particularly useful and versatile number system and digits, created and completed early in the Renaissance era, and thus we are capable of observing this quantitative phenomenon told by Newcomb and Benford centuries later and formulated in the context of numbers and digits, and that no one could have possibly expressed or stated such quantitative law in the Middle Ages before digits were invented? Would Roman Numerals show any meaningful pattern for large data sets, or would it be necessary for quantities occurring in the Roman Empire to be first converted to our modern positional number system before any pattern can be observed? Would Native Americans using the Mayan Number System be able to arrive at some sort of Benford's Law or its equivalent? Surely, a statement about relative quantities would be more general and universal and thus preferable to a statement about relative digital occurrences. Finally, the possibility of directly deducing  $\text{LOG}(1 + 1/d)$  statement about digital proportion from some more general statement about quantities seems like a tantalizing goal, one worthy of being investigated.

In the back of the mind of every student of Benford's Law there is a vague sense that there is more to the phenomenon than digits, and at times even the concepts of digits, numbers, and quantities are blurred. The most prominent characteristic of the generic Benford  $k/x$  distribution is its consistent density fall, and that no matter what range it is defined over, small quantities are more numerous than big ones. Such state of affairs in the discipline is rather unsettling; it calls for a revolution in the conceptual understanding of whole phenomenon of Benford's Law, putting it firmly on a quantitative basis, and especially so for its physical manifestation.

## BENFORD'S LAW IN ITS PUREST FORM

---

---

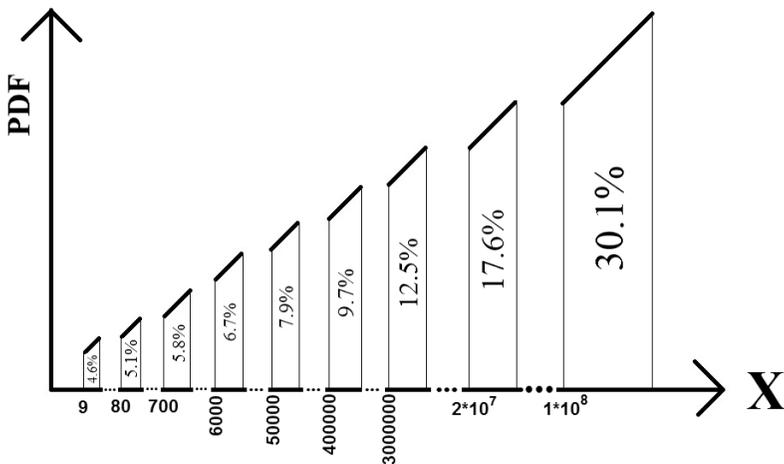
In Chapters 72, 73, and 74, the focus has been shifted to the properties of density curves of logarithmic data. It was shown that such density curves must be falling overall in the aggregate, having a long tail to the right, and that the fall comes with a particular overall sharpness, all of which must also be well-coordinated between IPOT points. Hence, the agenda should be turned around perhaps, from looking at the way digits occur within typical numbers of everyday data to looking at the typical forms of densities of typical pieces of everyday data, and the corresponding proposal is that the leading-digit phenomenon is very much a phenomenon of the propensity of everyday (positive) data to occur as lop-sided densities with tails to the right, falling off (or even rising temporarily) over the relevant intervals bounded by integral powers of ten with a variety of sharpness or rapidity, yet all having that same aggregate rate of fall over the entire range. Such a particular fall results in that ubiquitous  $\text{LOG}(1 + 1/d)$  leading-digits distribution.

Alternatively stated: it is not possible to separate leading-digits distribution from the nature of the density curve itself. Leading-digit distributions and density curves are not independent of each other and any statement about leading-digits configuration mathematically implies a restriction on the shape and range of the density curve. Surely, a statement about digits configuration cannot uniquely determine the density curve; there are in principle infinitely many densities satisfying  $\text{LOG}(1 + 1/d)$ ; yet the law does restrict the curve to a particular type and shape over a defined range. Consequently, this vista yields yet another essential perspective in Benford's Law regarding relative quantities, since a falling density curve to the right implies that small values are numerous and big values are few. Hence, as Benford's Law states a preference for lower digits, it also by implication states a preference for lower quantities as well, namely that 'small is beautiful'.

Would it be possible to detach Benford's Law as a statement about digits from a law about relative quantities? Can data be constructed in such a way that it obeys Benford's Law digitally while simultaneously possessing numerous big values and very few small values, namely an inverse quantitative configuration?

One straightforward approach is to focus solely on first-order digits, and position area under the curve pertaining to where digit 1 leads farthest to the right where the biggest values reside; position area where digit 2 leads just to the left of it; and position area where digit 9 leads on the leftmost section where the smallest values reside. Figure 7.1 depicts such a possibility where high quantities are always more numerous than low quantities, and where the big happened to be consistently more beautiful than the small. In the construction of the density curve, care was taken to insure that density is consistently rising throughout the entire range so that the distribution represents a truly inverted quantitative configuration as opposed to typical logarithmic configurations such as  $k/x$  and Lognormal curves where density is (almost) always falling. Areas were carefully constructed in such a way as to insure that each block corresponds exactly to  $\text{LOG}(1+1/d)$  according to what digit  $d$  dominates first place on the the x-axis below it.

What is the price that we are forced to pay for such an odd achievement? In fact there are two calamities stemming from this construction. The first is the profound and unnatural discontinuity and separation of the nine sub-ranges where density is defined over, which is extremely rare in real-life data. The second mishap is the severe rebellions spreading among all higher orders, not only refusing to obey the law of Benford, but also in a spiteful manner are reversing the legal



**Figure 7.1** A Digital Law Purely of the 1st Order — Full Quantitative Reversal

preference for low digits and are purposely and vindictively supporting high digits here, as if avenging that quantitative reversal.

Could another construction be made where all higher orders follow the law obediently and the general form of Benford's Law is observed? The way to go around the issue of higher orders is to utilize mini  $k/x$  distributions over each of the nine sections separately, as shown in Fig. 7.2 (A). It is necessary here to have each first digit go over a whole second-order cycle, thus density of first-order-digit 9 is defined over  $(9.0, 10.0)$ , and density of first-order-digit 8 is defined over  $(80, 90)$ , and so forth. To find the value of  $k$  for the section of, say, first digit 1, we solve for  $k$  in the expression  $\int k/x \, dx = 0.301$  over  $[1*10^8, 2*10^8]$ , leading to the expression  $k*[\ln(2*10^8) - \ln(1*10^8)] = 0.301$ , and thus  $k = 0.434295$ . In general this leads to  $k = [\log(d+1) - \log(d)]/[\ln(d+1) - \ln(d)]$  which then reduces to  $k = 1/\ln(10)$  or simply  $k = 0.434295$  for all the digits, implying that all nine disjoint curves in Fig. 7.2 (A) actually belong to the single curve  $0.434295/x$  over  $[9, 2*10^8]$ , albeit with stops and cuts. Hence in Fig. 7.2 (A), Benford's Law is obeyed in the general sense; all higher orders are distributed according to the law; mantissa is uniform; and to a certain degree big values are much more numerous than small values. Yet, within each of the nine sections, the small is still beautiful, and big values are outnumbered by small values consistently. Overall though, the general picture is that we have many more values of very big quantities, such as those residing in  $(1*10^8, 2*10^8)$ , less of smaller values residing in  $(2*10^7, 3*10^7)$ , and by far the fewest number of values of the smallest quantities residing in  $(9, 10)$ .

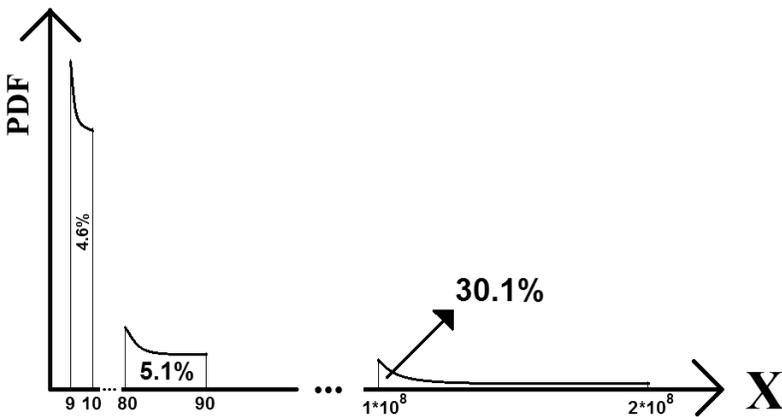


Figure 7.2 (A) General Benford Density —Yet the Big Overtake the Small in General

Yet, in spite of all our intense but misguided effort in combating the quantitative aspect of the law, the mode (highest density point) firmly resides in the smallest of all possible values, namely in 9!

Figure 7.2 (B) depicts a simpler and more straightforward arrangement yielding such partial quantitative reversal where the general form of Benford’s Law is exactly obeyed with all higher orders considered. Here distinct generic-Benford densities of the form  $k/x$  reside within each sub-interval standing between adjacent IPOT points, and where curves to the right are always higher than the curves to their left — in order to further emphasize that the big is overall more beautiful and numerous than the small. It must be noted that since each distinct area does not add up to unity, the classic relationship [discussed earlier in Chapter 60] of  $k = 1/\ln(\text{upper edge}/\text{lower edge})$  or  $k = 1/\ln(10)$  does not hold here. Rather we assign values for  $k_1, k_2, k_3$  for the three densities in an arbitrary manner to insure quantitative reversal. In Fig. 7.2 (B),  $k_1/x$  is defined over  $(1, 10)$ ,  $k_2/x$  is defined over  $(10, 100)$ , and  $k_3/x$  is defined over  $(100, 1000)$ . Insuring that total area adds up to unity requires that:

$$\begin{aligned}
 1 &= \int k_1/x \, dx \text{ over } (1, 10) + \int k_2/x \, dx \text{ over } (10, 100) + \int k_3/x \, dx \text{ over } (100, 1000) \\
 1 &= k_1[\ln(10) - \ln(1)] + k_2[\ln(100) - \ln(10)] + k_3[\ln(1000) - \ln(100)] \\
 1 &= k_1[\ln(10/1)] + k_2[\ln(100/10)] + k_3[\ln(1000/100)] \\
 1 &= k_1*\ln(10) + k_2*\ln(10) + k_3*\ln(10) \\
 1/\ln(10) &= k_1 + k_2 + k_3
 \end{aligned}$$

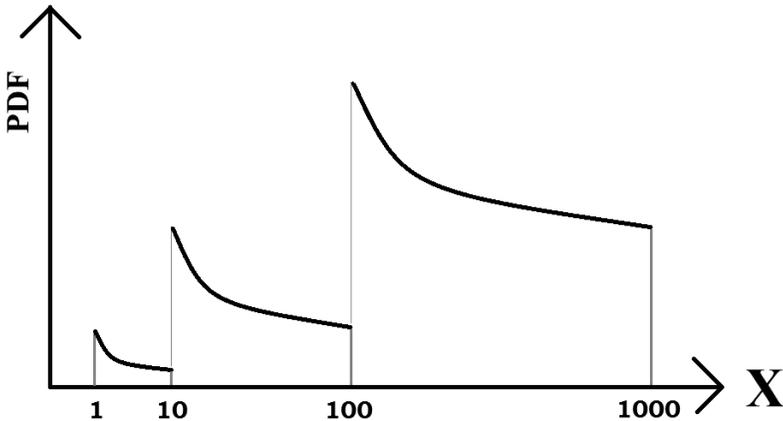


Figure 7.2 (B) Discontinuous  $k/x$  Benford Densities with Partial Quantitative Reversal

Within each of the three sub-intervals of Fig. 7.2 (B) Benford's Law is exactly obeyed and mantissa is uniform; in spite of the fact that  $k_i \neq 1/\ln(10)$  as in all classic  $k/x$  distributions standing between adjacent IPOT. The aggregate of the mixing of 3 such data sets surely obeys Benford's Law as well since it is composed of 3 perfectly logarithmic parts. Admittedly, within each sub-interval the small is still more beautiful than the big. It is only when the data set is considered in its entirety that we obtained some quantitative reversal here overall. The price paid for such partial quantitative reversal is the severe discontinuity at 10 and at 100. Such severe discontinuous data arrangement is extremely rare in real-life data sets. In almost all real-life data sets, the tail of the section on (1, 10) connects (at 10) with the head of the adjacent section on (10, 100); and the tail of the section on (10, 100) connects (at 100) with the head of the adjacent section on (100, 1000), and so forth as depicted in Figure 7.2 (C), insuring that the small is consistently more beautiful than the big. What Fig. 7.2 (C) shows is that the confluence of continuity and Benford-ness guarantees (an almost) consistent quantitative decline everywhere. Admittedly, Fig. 7.2 (C) relates only to the unique and rather rare deterministic flavor of the law, yet when typical real-life data sets of the random flavor are considered, very similar densities are found which fall almost everywhere except at the very beginning for very low values [where density is sharply but very briefly climbing upwards, as in all Lognormal distributions with high shape parameter]. Conceptually it is crucially important to acknowledge that for data on any segment between two adjacent IPOT points such as say (10, 100)

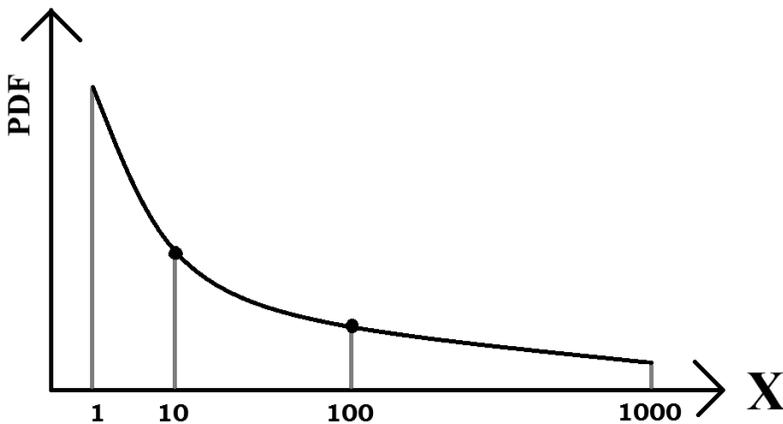


Figure 7.2 (C) The Confluence of Continuity and Benford Yields Quantitative Decline

there is no way to distinguish between digital preference for low digits and quantitative preference for low values, as one aspect implies the other aspect of the phenomenon.

In conclusion, it is not possible to strictly separate Benford's Law about first and all higher-order digits from its implied law about (all the) quantities, although such separation can be done in a general sense for groups of quantities. Admittedly, if one is willing to break down the last remnant of continuity and go along exclusively with discrete values, then for the careful collection of finite data points along the x-axis, beginning sparsely on the left and gradually increasing in frequency and concentration on the right, where all orders obey the law approximately, the big could still be consistently more beautiful than the small.

Yet, data such as the ones depicted in Figs. 7.1, 7.2(A), and 7.2(B) have to be artificially concocted in the imaginative mind of the data analyst; they appear farfetched and removed from real-life typical data sets and are almost never empirically encountered. Almost all data sets show continuity and compactness, and are rarely broken down into separate and disjoint intervals. All this is a strong confirmation that Benford's Law is very much a law about quantities as much as it is about digits, albeit indirectly so.

## NUMBER SYSTEM INVARIANCE PRINCIPLE

---



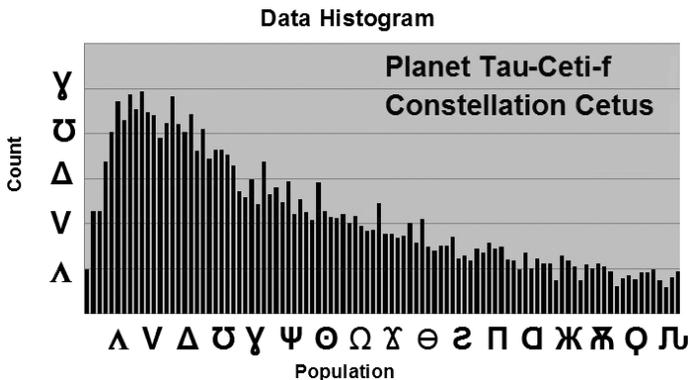
---

Would it be possible to separate Benford’s Law from our number system itself altogether or from any number system for that matter? Could the law be totally independent of any artificially man-made number system with digits? The answer is in the affirmative, and in order to demonstrate this, the data on U.S. population centers depicted in the histogram of Fig. 4.50 shall be re-considered. As discussed earlier in Chapter 72, the notable feature of that population histogram is its diminishing density on the right, having numerous small quantities and few big ones, and this attribute (pertaining to **any** logarithmic data set) has its basis in theoretical reasoning as well as in consistent empirical findings. It is scale-invariant as well as base-invariant. Moreover, and by far more significant, it is number-system-invariant, reality floating out there in the universe, independent of whatever we do, calculate, count, invent, or define. To the extent that Benford’s Law is a quantitative statement as well, this affords the law universality, rendering it a physical and scientific law.

On planet Tau-Ceti-f (constellation Cetus), in the wake of horrific and brutal wars over resources, excessive exploitation, rampant slavery, and suffocating monopolies in all industries and commerce, the emerging benevolent dictator decrees the abolition and forbidding of **any** number system, which is perceived as being the principle instigator and cause of greed, facilitating the easy counting and imagining of quantities of money and lacking an intrinsic upper limit. Frequent cries of “numbers are at the root of all evil” are heard in all quarters. Endowed with superb memory and savoring their peace, Tau Cetians now assign a unique and singular symbol (‘digit’) to each and every numerical quantity, up to 37,859 — their legal limit — and known symbolically as **Λ**. Severe social castigation and prosecution befall all those who merely mention a quantity larger than **Λ**, even in non-monetary contexts. All children easily master the obligatory 37,859 numerical symbols {**Δ V Δ Ξ Υ Ψ Θ Ω Σ Θ Σ Π Δ Ж Ж ϕ Λ** and so forth to **Λ**}, but they do have trouble adding and multiplying. For an Earthling, the symbol **Λ** signifies 1, the symbol **Δ** signifies 3, **Ж** signifies 15, and **Λ** signifies 37,859. To the amazement of the chief statistician at the Galactic Federation Census Headquarters, it was

discovered that the biggest domain in Tau-Ceti-f, the United States of Numerica, has by some very rare coincidence the same (physical) population configuration as that of USA of Earth. Deeply disdaining both planets for their past massacres and genocides (both are still on the Galactic Watch List), the statistician nonetheless holds his contempt in check and muses over the thought that their petty statisticians should stare at two identical histograms in spite of their profound difference in how they relate to quantities. While Earth excels in mathematics, utilizing one of the most efficient number systems in the whole of the Galaxy, Tau-Ceti-f in sharp contrast demonstrates its numerical backwardness, where the very concept of a base such as 10, e, or 2, does not exist. Yet both planets come up with the same visual result of a histogram that is constructed and based exclusively on quantities! Figure 7.3 depicts the histogram of population data on Tau-Ceti-f for US of Numerica, which is identical to the one in Fig. 4.50, except for the designations of quantities on the axes. For better visualization, the horizontal scale for Tau-Ceti-f is misaligned as it shows only 17 numerical symbols (up to symbol  $\text{J}$ ), instead of the necessary full set of 1,400 numerical symbols, for lack of space. The vertical scale is misaligned as well, showing only five numerical symbols instead of the necessary full set of 300 symbols, for lack of space.

A moment's thought would convince anyone that these two histograms really are visually identical, just as the chief galactic statistician believes them to be, if one ignores all the strange or familiar digits and symbols around them. This fact can be easily verified if we note that both the X and Y axes represent primitive count, which does not necessitate any number system. The X axis represents population count value of a city or a town while the Y axis represents the count of how many



**Figure 7.3** Identically Shaped Histogram on Tau-Ceti-f — Absent a Number System

cities and towns exist having such a population count. The correspondence in histogram's shape found here is actually more general than merely count histogram, as it applies universally to any histogram type, including those requiring a scale to measure continuous quantities. A scale change only causes the histogram to be either squeezed or stretched horizontally. After all, we express continuous quantities by **counting** how many units of the atomic scale fit the physical object in question, plus perhaps some fractional part. For example, when we state that a certain car model weighs 767.25 kilograms, we say that we can count within it the equivalent of 767 units of kilogram, plus a quarter of that kilogram, hence conceptually scale measurements also relate to pure counts.

Probability of the proportions of symbols on Tau-Ceti-f for the first and only order of all its 37,859 digits/symbols vary according to the specific data set in question, as they are exclusively data-driven (i.e. data-specific); there exists no law. No law whatsoever can be stated if society does not utilize repetitive cycles of sub-intervals similar to those on Earth standing between IPOT such as (1, 10), (10, 100), (100, 1000), and so forth. Each data set and its associated histogram is unique, and on Tau-Ceti-f there exist no digits to connect and unify diverse data sets via some universal digital proportion. Older books had extensive expositions on the currently prohibited topic of Zikuma's Law, namely uniformity of mantissa and the expression of  $\text{LOG}_9(1+1/\text{digit})$  for first-order proportions in the older system of base 9 (derived from having three fingers on each of the three arms, nine in total). These invaluable books were burnt in haste along with all of their mathematics and engineering books in order to preserve the fragile peace. On the other hand, everybody is encouraged to contemplate and revere their so-called 'Sacred Pacific Proportions', namely the yearly publications of a long vector of the specific probability proportions of all their 37,859 digits, calculated on the last minute of each year for the entirety of their data, as in the Aggregate Global Tau-Ceti-f Data Interpretation of the law.

Another crucial observation and the second moral of the story here is that there could be no meaningful law to contemplate if not for that repetition in the use of digits or symbols, namely a positional (place-value notation) number system or something similar to it perhaps. We here on Earth reapply the digits in ever widening magnitudes, and it is exactly this feature within our number system that facilitates a meaningful discussion about proportions of digits. For example, digit 8 means different things within different numbers, so that for 8 it means 8, for 5890 it means 800, for 8777111 it means eight million, and so forth. Such a feature is notably absent on that peaceful yet numerically backward Tau-Ceti-f planet.

## CARTESIAN COORDINATE SYSTEM IS NUMBER-SYSTEM-INVARIANT

---

It is quite astonishing to observe that the Cartesian coordinate system originated only about 350 years ago in the 17th century by René Descartes, the intellectual giant of the late Renaissance era. It is exceedingly hard to imagine life without this essential tool! Figure 7.4 depicts the generic form of the plane. Absent the concept of a horizontal X line perpendicularly superimposed on a vertical Y line to constitute the Cartesian Plane, very little can be done in scientific, mathematical, and engineering endeavors. Statistical data histograms and density distributions are constituted on nothing else but the Cartesian Plane! It is not known whether Descartes himself had in mind applying his idea in analyzing and representing measured data as well, but since this book is full of them and since only his plane facilitated the intuition, insight and visualization here, the author is deeply grateful to him.

It must be noted that Descartes never insisted on mixing our efficient number system with his plane. The conceptual insight of the Cartesian Plane does not necessitate our number system, nor any other number system for that matter. It can be thought of as a purely geometrical construction where distance and the primitive counting of unitary squares play the leading roles. Each unity of distance on the axes is duplicated and a count to the origin signifies quantity or position. Does it matter whether we call the tenth unit 10,  $\Theta$ , or 1010 in the binary number system? Certainly not! Therefore, any pictorial collection of points or continuous curve on the Cartesian Plane representing quantities is base-invariant and even number-system-invariant — except for the designation of those quantities below the X axis and to the left of the Y axis. This is why the two histograms of data on populations appear identical on Earth and on Tau-Ceti-f. The only time the two planets differ is when Earthlings represent data utilizing their fancy log-scale coordinate, which often saves paper and facilitates the demonstration of long and complex charts. Had the population of any Egyptian, Mayan, Roman, or Persian province possessed the same configuration as that of USA in 2009, then surely their histogram would also look identical, except for those peculiar symbols below and

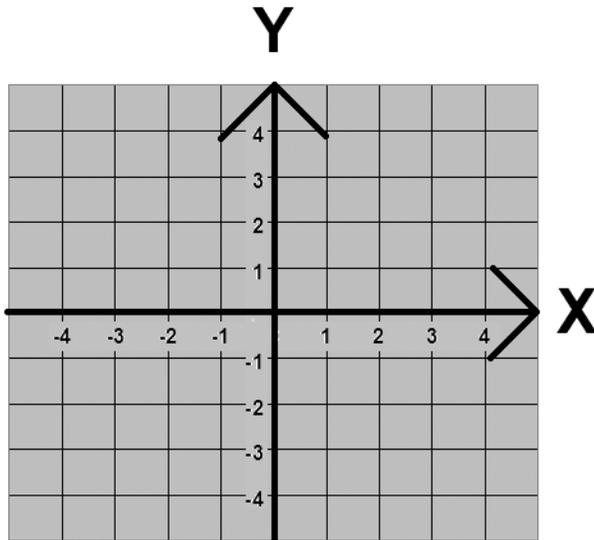


Figure 7.4 Cartesian Coordinate System

to the left, assuming someone innovative enough back then and well before René Descartes could have come up with such an unusual idea as a scatter plot. In any case, while data histograms and distribution densities turned out to be universal and identical in all civilizations regardless of the number system in use, related LOG histograms or densities are not base-invariant, and they do appear differently for each distinct base, although flatness and curviness are independent of base (implying that the distinction between the random and the deterministic flavors is base-invariant). For number systems that are not positional (i.e. without place-value notation), related log cannot even be constructed or defined and so comparison is not possible. Related log conjecture, for example, which was suggested and extensively explored in our decimal positional system, is applicable to civilizations applying other bases, but not for those applying non-positional ones or without any number system whatsoever such as the one on Tau-Ceti-f.

---

---

## PHYSICS IS NUMBER-SYSTEM-INVARIANT

---

---

When Isaac Newton proclaimed his second law of motion  $F = MA$  and the associated gravitational law  $F = GM_1M_2/R^2$  no one raised the slightest objection about the possibility that it may lack universality; that it may not hold in other base systems; or that a change in scales for length, mass, and time would cause havoc upon the law. Pre-revolutionary Tau Cetians with nine fingers also used to observe and write extensively about Newton's laws using their old base 9 number system just as Earthlings do now with their base 10 number system; and clearly Newton's laws are base-invariant as both sides of the equation go under identical and corresponding transformations. The fact that Tau Cetians use the Largoza as the unit for length, the Pezada as the unit for mass, and the Tempora as the unit for time, does not preclude their affirmation of law; they simply believe in a different value of  $G$  as the gravitational constant and everything falls into place. Post-revolutionary Tau Cetians are permitted to continue writing about and applying Newton's laws, but now all is done without any number system except by way of their cumbersome symbolic expressions of quantities; and that extremely large stars, planets, distances, and forces are prohibited from being analyzed or even mentioned lest greed return. Do they still observe Newton's laws? In other words, do Newton's laws depend on our number system, or any number system per se in order to be stated, or in order to be valid? The obvious answer is that the laws are number-system-invariant; Tau Cetians as well as Earthlings observe the same laws. Newton's statements are universal and quantitative laws, not numerical ones.

No one, not even Newton himself, has ever seen or experienced pure forces. Rather, we deduce or theorize them from the observational motion or acceleration of physical objects. Therefore his laws are really about the physically verifiable statement that  $GM_1M_2/R^2 = M_2A$ . Let us consider  $M_1$  as a much heavier object such as a star whose motion due to gravitational interaction is barely (if at all) noticeable, and  $M_2$  as a smaller object such as a planet whose motion is measured.

The term  $M_2$  cancels out; the term  $M_{\text{STAR}}$  is substituted for  $M_1$ , and his statement then reduces to:  $A_{\text{PLANET}} = GM_{\text{STAR}}/R^2$ . Physicists on Tau-Ceti-f plotting acceleration  $A$  versus distance  $R$  observe the histogram shown in Fig. 7.5, which is identical to the one prepared by Earth physicists except for symbols. Also the histogram in Fig. 7.6 of (planet) acceleration  $A$  versus (star) mass  $M$  on Tau-Ceti-f is identical to the one prepared on Earth, except for the different symbols and digits below and to the left.

The law  $GM_1M_2/R^2 = M_2A$  is a purely quantitative statement in nature. The Babylonian, Egyptian, Roman, and Mayan civilizations would all welcome and accept Newton's law. Newton equates quantities and lets you decide on the number system you like to count and express those quantities. He lets you divide and multiply however you like, but you have to understand the primitive concepts of multiplications, divisions, and equality. Had Roman Numerals been still in use in Europe around his time, there is no reason to think that young Newton wouldn't have been able to discover classical mechanics.

Consequently, if a good part of the phenomenon of Benford's Law is derived from the multiplicative forms of the equations in physics and so forth [as discussed in Chapter 90], then this pattern should be observable on all planets, and regardless of the particular number system locally in use, or absence thereof.

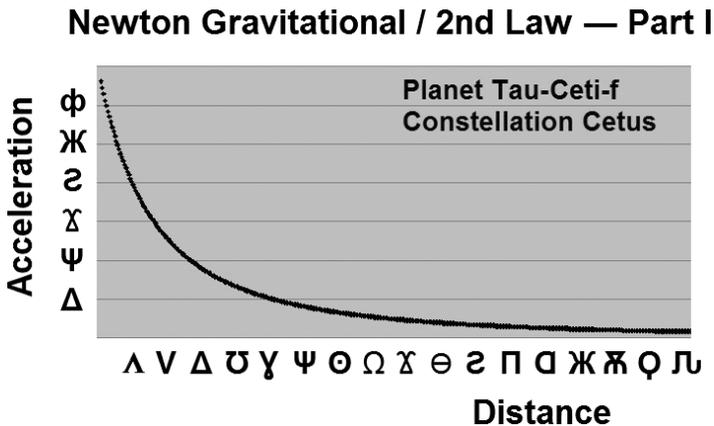
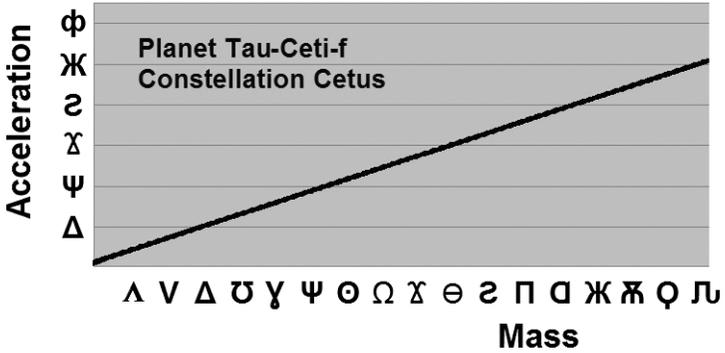


Figure 7.5 Law of Acceleration vs. Distance on Tau-Ceti-f — Absent Number System

**Newton Gravitational / 2nd Law — Part II**

**Figure 7.6** Law of Acceleration vs. Mass on Tau-Ceti-f — Absent Number System

## MULTIPLICATIVE CLT IS NUMBER-SYSTEM-INVARIANT

---

It was conjectured earlier that the physical manifestation of Benford's Law (single-issue data which cannot come under the protective umbrella of Hill's model) may be a partial consequence of the Multiplicative Central Limit Theorem (MCLT) which points to the Lognormal-like as the underlying distribution. Certainly, MCLT (or its partial application in conjunction with Related Log Conjecture) plays a major role in Benford's Law, and Hamming's remark that 'Benford's Law seems to appear out of nowhere' when variables were repeatedly multiplied is but one example of its importance and relevance in the field. Thus any demonstration that MCLT transcends number systems and digits would in and of itself be decisive for any such invariance conclusion. But clearly, the type of data structure that MCLT builds is purely quantitative in nature; it is number-system-invariant, necessitating neither digits nor numbers.

Figure 5.2 in Chapter 90 on the distribution of the product of four dice depicts just one example of the typical and generic histograms obtained as a result of repeated multiplications associated with MCLT. As the shrewd and manipulative casino owner already knows, it is skewed to the right having numerous small quantities and very few big ones. How would the histogram of such product of four dice look like on post-revolutionary Tau-Ceti-f planet? There is no data available on such topics whatsoever, not even at the Galactic Federation Census Headquarters, since casinos and gambling are strictly forbidden now on that peaceful planet in order to further curb greed, but one can easily imagine and construct hypothetical histogram for them, as depicted in Fig. 7.7. As discussed earlier, the use of a different base in positional number systems, the adoption of any other numerical system, and even the absence of a number system altogether do not alter histograms in any way. Histograms are planet-invariant, as is the universality of the motto "small is beautiful".

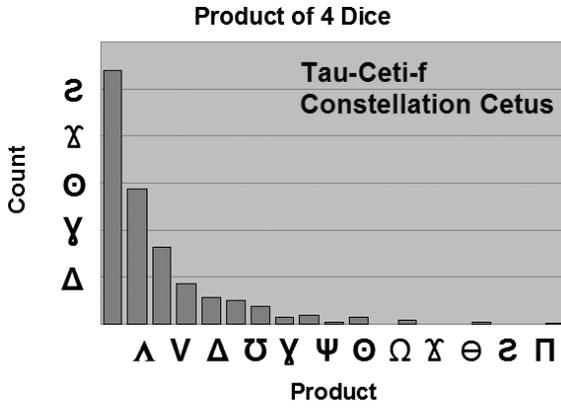


Figure 7.7 Histogram of Products of Four Dice on Tau-Ceti-f Absent a Number System

## GREEK PARABLE AND CHAINS ARE NUMBER-SYSTEM-INVARIANT

---

The Greek Parable demonstrates how conversations about quantities in a primitive society lead to something quite similar to Benford's Law. Its (almost) universal nature was noted earlier in Chapter 50, as seen in its total independence of any scale or unit of measurement, as well as its total independence of any choice of a base within a number system. Far more significantly, the parable does not even necessitate or assume any number system in and of itself! The moral of that story is quantitative in essence, not numerical. The 'numbers' or 'digits' 1 through 9 might as well been some Egyptian or Mayan symbols, and still the whole mathematical edifice does not collapse in the least. Since 9 was the upper limit there, there was no need to express quantities via our number system at all, nor via **any** number system for that matter. For example 537 necessitates the use of our number system, and it is expressed as  $5*100 + 3*10 + 7*1$ , but the 'number' 7 say does not need to be expressed by way of a number system. It stands on its own independently as a symbol of the quantity seven; and in fact it also serves as one brick in the edifice of our number system itself – as a digit. More vividly put: for an antiquated Tau Cetian society living under the same conditions described in the Greek Parable with only nine objects and only nine numerical symbols  $\{\Lambda, V, \Delta, \Upsilon, \Psi, \Theta, \Omega, \Xi\}$ , the same exact resultant distribution follows!

By extension, the results of the simple averaging schemes, more complex averaging schemes, as well as infinite averaging schemes, are also number-system-invariant. Does the histogram-like of the Greek Parable such as the one seen in Fig. 4.57 change when Roman Numerals or Tau Cetian symbols are in use? Surely not! Does extending upper limit to a quantity well over nine alter the quantitative arrangement of the scheme? Surely not! There is nothing intrinsically special in the quantity nine. Hence histograms of all averaging schemes are number-system-invariant and quantitative in nature. Since chains of uniform distributions and averaging schemes mirror each other, the same invariance conclusion is applicable to those chains as well.

## PHYSICAL REALITY VERSUS DIGITAL PERCEPTION

---

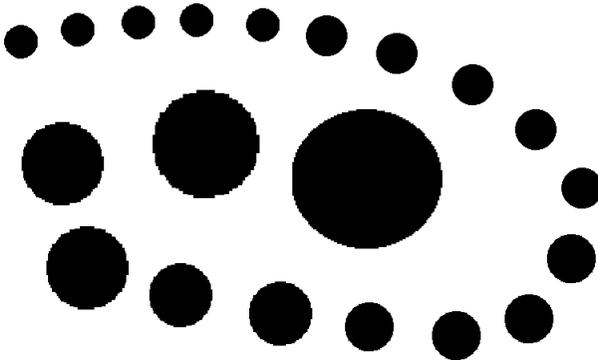
In light of such vast evidence that the logarithmic is highly prevalent in the physical world, such as in data on pulsar rotation, earthquake, river flow, chemical molar mass, population centers, and so forth, two radically different interpretations of the phenomenon are given. The first interpretation is that this is truly a physical phenomenon existing independently of us and our way of recording data; that it is a physical law of nature that can be measured and detected by ways other than our digital perceptions. The second interpretation claims that this digital pattern found in physical data is simply due to our own peculiar way of counting values by way of their digital representations, and therefore the phenomenon has no independent physical existence outside our digital perception, that it is merely a law regarding our own self-constructed number system and digits. As an analogy for the second interpretation, a child wearing red eyeglasses may believe that every physical object in the world is red. A distorted telescope with a defective lens may convince the observer that all things are elongated. The extremely naïve statistician working on multiple physical data sets in a windowless room full of computers and living in a society which utilizes base 3 with digital symbols  $\{\Lambda, \mathbf{V}, \Delta\}$  may ludicrously believe that all physical things relate to or are essentially triangular by nature. Healthy skepticism in the spirit of David Hume should not be considered as some mere philosophical entertainment, but rather as an essential tool that leads to concrete results and better insight.

In this section it shall be shown that the first interpretation is the correct one, that physical data sets do possess a certain pattern independently of us, our number system and our digits, and that the pattern is innate and deserving of the label 'a law of nature'. Yet, for such property to be considered universal and independent of societal number system, it must be stated in terms of quantities rather than in terms of digits.

## PATTERNS IN PHYSICAL DATA TRANSCEND NUMBER SYSTEMS AND DIGITS

---

To further demonstrate the number-system-invariance property of Benford's Law and to lend the (above) first interpretation decisive support, a snapshot of 20 imaginary exoplanets is shown in Fig. 7.8. The figure depicts what can be easily visualized, namely their sizes, which are assumed to be Benford. The measure of size here may refer to the two-dimensional area (as in telescopic observations), which is of course derived from or directly related to true sizes such as diameter, radii, or volume in a one-to-one relationship. Clearly this feature of the data having numerous small ones and few large ones in a very certain and peculiar manner (the logarithmic manner) can be seen and understood without any number system or digits whatsoever. Those 20 planets appear identical on all planets and civilizations, regardless of their number systems or absence thereof. This snapshot of exoplanets represents something very generic and basic — the feature lying at the very heart of Benford's Law. Similar pictures and snapshots are seen for all other physical logarithmic data sets, be they rivers, populations, and so forth, and it's rather impossible to argue that this is all about digits or even about numbers. Rather, it's all about relative quantities and relative sizes! All snapshots about logarithmic physical data sets give out the same message: 'small is beautiful'.



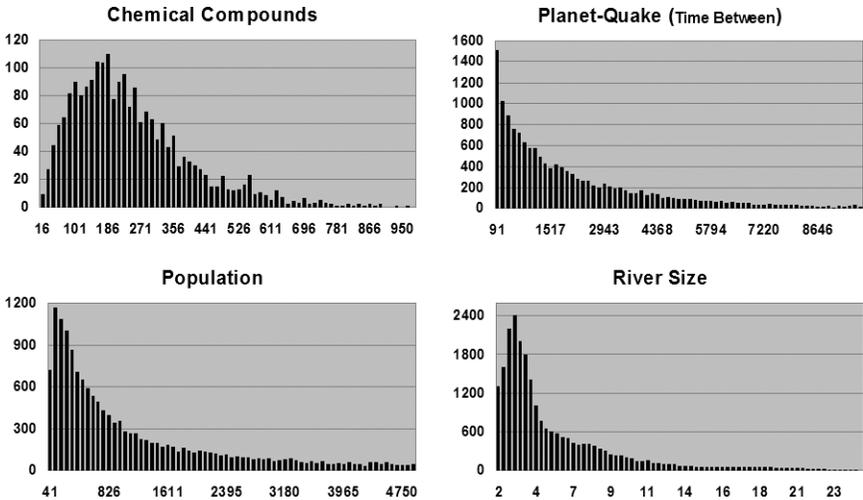
**Figure 7.8** Planets' Logarithmic Sizes Are Visualized Independently of Numbers

## COMMON THREAD GOING THROUGH MULTIPLE PHYSICAL DATA SETS

---

Galactic Year is defined by the Federation as the successful completion of a full rotation of the Milky Way relative to the surrounding Local Group of 54 Galaxies. Celebrations of the beginning of each new Galactic Year are encouraged and promoted by the Federation in order to increase harmony and sense of unity and belonging among the various planetary factions. In accordance with its policy of investing in galactic education, the Federation has a long tradition of declaring such auspicious moments 'Educational Dawn for the New Galactic Year'. Its chief statistician was assigned the difficult task of spreading knowledge about Benford's (Zikuma's) Law throughout the galaxy. While his contact with planet Earth went fairly smoothly, his communication with planet Tau-Ceti-f proved the most difficult and contentious. To simplify matters, he presented to the chief planetary statisticians of both planets physical data (containing only positive values) on the well-known planet Clarikia, easily observable by almost all other planets. The four data sets involved relate to: (1) molar mass of the most frequently used chemical compounds on Clarikia, (2) time between Clarikiaquakes, (3) population centers, and (4) river sizes. Figure 7.9 depicts the four histograms hurriedly sent (almost instantaneously) to the two Planetary Statistical Headquarters via the expensive Antirelativistic Rapid Communication Technology System a.k.a. ARCTS (still passionately detested and avoided by theoretical physicists to this day, but beloved and operated by enthusiastic engineers who do not really understand what they are doing). Time pressure and the urgency of the matter have caused the chief galactic statistician to forget to translate the first set sent to Earth accompanied with their digits into Tau Cetian numerical symbols. Knowing that he did a sloppy job he nonetheless found sufficient solace in the fact that the main body of the histograms is universal and common to both planets and thus correct.

Earth's prompt reply came immediately, acknowledging that the sets of the proportions of the first nine digits of the four histograms came out almost the same, closely fitting  $\text{LOG}(1 + 1/d)$ , and that this pattern is long known and



**Figure 7.9** Four Histograms on Clarikian Physical Data — Universally Observed

applied for millennia, catching many a thief. The reply from Tau-Ceti-f, however, was delayed by the uproar and rage the message from the Federation caused there due to the inclusion of these offensive numerical digits in the message, in clear violation of Galactic Constitution clause XLVII guaranteeing respect for local traditions and customs. In a hastily arranged meeting with the dictator, the generals donning ridiculous medals and silly-looking stars and conveying supreme self-importance argued that the event constitutes *casus belli* and urged to war with Earth (which had nothing to do with the Federation's message), hoping to preserve their dwindling status and salaries. But the dictator would have none of this, and referred the whole matter to the chief planetary statistician. While adamantly refusing to superimpose those forbidden digits onto the four histograms as a matter of principle, the chief Tau Cetian statistician nonetheless promised the Federation to look seriously into the matter and to come up with a statistical measure on all four histograms that does not involve any offensive digits, a measure that would be approximately constant across all four data sets, and thus constituting a general law.

How should Tau Cetians go about finding a quantitative pattern in the data sets given that digits are not to be allowed? The Federation sent a second message apologizing for the offense and providing the actual values (this time in Tau Cetian symbols) of all four data sets should this more detailed information be necessary

in the construction of the statistical measure. Fortunately for all involved the highest value in all four data sets was still less than  $\mathbb{A}$ , the Tau Cetian legal limit of 37,859.

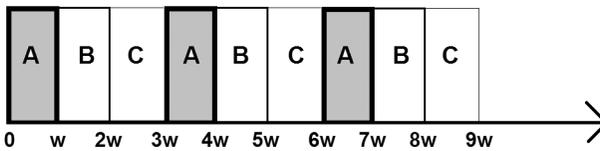
In the message, the chief galactic statistician also personally assured the Tau Cetian statistician that there is a verifiable common measure of fall in those four Clarikian histograms in spite of the very distinct way they visually appear, and that that measure would also be applicable in most other data sets on Tau-Ceti-f. The very fact that the pictorial or visual aspect of these four histograms is universally observed on all planets throughout the galaxy necessitates that they could all come up with a universal, primitive, and very basic statistical measure agreed by all, namely being planet-invariant and totally independent of any artificially invented number system unique to the local cultural and mathematical background of any given planet. In other words, that a singular quantitative Benford's Law statement should be true (identical) for all observers. Had these four data sets and histograms been about abstract numbers provided by Clarikian accountants and economists and such, then one may argue that a law or regularity within these four charts may uniquely apply solely to them due to their unique number system, and that these values were created within the context of their peculiar number system. For example, data sets such as house value appraisals, enormous quantities of money supply generated abstractly on the keyboard periodically by their central bank, arbitrary penalty amounts ruled by their judges against offending parties, and so forth, are all examples of 'quantities' generated in the context of a number system, contemplated and utilized in the minds of their appraisal agents, central bankers and judges. But since the four histograms above represent physical data not generated by way of any number system, but rather handmade by Mother Nature herself (who never learnt of numbers, digits, and symbols) and are universally observed throughout the galaxy, laws and regularities here must also be made universal and number-system-invariant.

## CASTING A REPETITIVE BIN SYSTEM TO MEASURE FALL IN HISTOGRAM

---

It occurred to the chief Tau Cetian statistician that in order to measure overall relative occurrences of quantities or fall in histograms, one could simply cast a long system of repetitive bins along the x-axis and record the relative values falling within each bin. The idea emanates from the fact that, in general, flat histograms yield data with as many big quantities as small or medium ones. Rising histograms yield data with mostly big quantities and fewer small ones. Falling histograms yield data with numerous small quantities and fewer big ones. Hence, by constantly examining local relative fall within bins and then aggregating all results, we construct a singular measure of relative quantities (i.e. overall fall or rise in histogram). This scheme is somewhat akin to Benford's Law which uses nine digital bins, equally spaced, within each sub-interval  $[0.001, 0.01)$ ,  $[0.01, 0.1)$ ,  $[0.1, 1)$ ,  $[1, 10)$ ,  $[10, 100)$ , and so forth (when 10 is the base). The above incomplete set of these five Benford's digital/bin cycles of course omits infinitely many other groups of bins above 100 and below 0.001. The scheme illustrates the concept of viewing Benford's Law as a bin system. Within  $[1, 10)$  for example, Benford had set a net of nine equally spaced bins,  $[1, 2)$ ,  $[2, 3)$ , . . . ,  $[8, 9)$ ,  $[9, 10)$  to trap data falling inside each, and record the result. Benford casts similar nine-bin groups on other sections of the x-axis of course, but each section is of different width, expanding by an inflation factor (designated as  $F$ ) of 10, namely the base. Benford then finally aggregates all the data falling on all bins designated 'digit 1', 'digit 2', etc. to arrive at something close to  $\text{LOG}(1 + 1/d)$ .

Figure 7.10 depicts a small part of a flat three-bin system devised initially by the Tau Cetian statistician as a trial test. We define **D** as the number of bins (equivalent to the number of 'first-order digits', or 9 for Earth), although the letter  $D$  does not strictly stand for the initial letter in the word 'Digit', but rather for the more general concept of bins. The term 'flat' signifies that the width is not expanding but rather is made to be a constant (i.e. that inflation factor  $F$  is 1). While Fig. 7.10 shows only three cycles of three-bin batch each, the scheme



**Figure 7.10** Non-Expanding Flat Three-Bin System Cast onto the X-axis

actually continues to the right indefinitely, or terminates sufficiently far on the right where all data points can be captured. The first cycle starts exactly from the origin 0. The width  $W$  is made intentionally to be neither too small nor too big. Width that is too small compared to the outlay of the data is in effect too refined to show any meaningful differentiation in data concentration (especially when for example no more than one value falls within each bin). Width that is too big compared to the outlay of the data is in effect too crude, as it captures a significant portion of the data within the first bins of the first cycle, distorting its message about relative concentration. Width that is made so big as to capture the entire data within the first bin of the first cycle renders the whole scheme meaningless.

Results badly disappointed the statistician. No law was observed. When width was made too small all three bins obtained almost equal portions, namely bin equality (for all four Clarikian data sets as well as for all local data on Tau-Ceti-f). It was only when bin width was made to be large as compared with the overall spread of the data set that bin skewness finally emerged, and in favor of 'low bins' as expected (with skewness correlating directly and positively with the size of the width). It was correctly reasoned that this skewness is perfectly consistent with the visually sensed overall fall in those four Clarikian histograms. Yet, since skewness (results) varied according to the value of the width  $W$ , no consistent law could have been stated. In other words, as width was made to be sufficiently large so as to escape bin equality, bin proportion then became simply a function of the width, as opposed to a steady and consistent law independent of width.

A major breakthrough occurred when the chief statistician acknowledged that so much of real-life data is the result of multiplication processes, and the prevalence of MCLT-driven data sets as well as exponential growth and decay. It was noted that values are 'stretched out' and 'expanded' along the x-axis 'rapidly' and 'forcefully' in a multiplicative manner much as was seen in the example of product of four dice earlier, and also as can be clearly noticed within the basic  $9 \times 9$  multiplication table (considered as a data set) that pre-revolutionary children

had to memorize in elementary schools generations earlier. The idea was then to utilize an expanding bin scheme, letting the width of the bin expand multiplicatively by an **inflation factor F**. If the data itself expands multiplicatively, then its measuring net of bins should also expand multiplicatively just the same, in order to 'keep up with it' and to be able to properly record it. Old Zikuma's Law must have also played a pivotal role in providing another clue and inspiration in this development, as it uses exactly such a scheme, with eight digital bins expanding by a factor of nine (their old base) on each cycle, but the prudent and wise statistician wouldn't admit any of this out of fear. On Earth for example, (1, 10) has a total width of 9 which is equally divided and allocated to each of the nine first-order digital bins. The width of the next cycle, namely of (10, 100) is ten times as wide at 90, and so forth, meaning that inflation factor F is 10. In general, the environment in which Benford's Law operates is one where digital bins are expanding by the value of the base, namely that inflation factor  $F = \text{BASE}$ . Since the number of first digits (D) in any number system is always  $(\text{BASE} - 1)$ , as in 9 first digits within number system base 10 applied on Earth, therefore **Benford's Law operates in the environment  $F = D + 1$** .

It is not only necessary to break out of Benford's narrow digital and numerical mold and to view instead the whole methodology of Benford's Law purely as arbitrarily constructed bins collecting data falling within, but it is also necessary to abandon the rigidity of Benford's Law in terms of D and F value combinations, and allow for any values of D and F in a new spirit of flexibility, insisting only on the constraints that  $F > 1$  to avoid the flat meaningless bin-equality, and  $D > 1$  to avoid a lone bin capturing the entire data set by default, rendering the whole scheme meaningless and the result a tautology.

Two expanding bin schemes were actually performed by the Tau Cetian statistician. **Scheme A** has 4 bins with an inflation factor of 8. **Scheme B** has 7 bins with an inflation factor of 3. Naturally it was decided to perform both schemes A and B starting from 0, and consequently it was deemed necessary to make bin width W start quite narrowly at 0.0008 in order to cast a refined net of bins applicable to data falling on (0, 1) as well. The value 0.0008 refers to the width of any bin within the first cycle, while on the second cycle bin widths are either  $8 * 0.0008$  for Scheme A, or  $3 * 0.0008$  for scheme B, and in general  $F * W$ . On the third cycle, bin width is  $F * F * W$ , and so forth. This aspect of the schemes represents another similarity to Benford's Law which operates under infinitely refined partitions on the left near the origin, and very crude and wide partitions on the right. We shall

use the following notations: D for the number of bins ('first digits'); F for the inflation factor ('base'); W for the initial width of a bin in the first set (first cycle) on the left; and S for the starting point of the whole scheme on the x-axis, namely the location of the left corner of the first ranked bin in the first cycle (which is usually assigned the value 0 in order to avoid missing data near the origin). These two schemes delighted the statistician, because all four Clarikian data sets sent to him by the Federation gave nearly the same results (i.e. bin proportions). The average of the proportion results of the four data sets within the context of Scheme A constituted a relative quantity law in the eye of the statistician, and another such law within the context of Scheme B. When other local Tau Cetians data sets were examined, their bin proportions also closely matched the two newly discovered bin laws, further increasing the enthusiasm of the statistician and lending more credence to the whole project. The summary of the setups of the two schemes and their results (the newly proclaimed laws, i.e. the averages from the four data sets) are as follows:

Scheme A:

$$D = 4 \quad F = 8 \quad S = 0 \quad W = 0.0008$$

$$\text{Proportions} = \{48.8\%, 23.5\%, 15.7\%, 12.0\%\}$$

Scheme B:

$$D = 7 \quad F = 3 \quad S = 0 \quad W = 0.0008$$

$$\text{Proportions} = \{22.6\%, 18.1\%, 15.5\%, 13.2\%, 11.1\%, 10.3\%, 9.2\%\}$$

Figure 7.11 depicts a small segment of Scheme A (just the leftmost part of the entire system) where the expansion factor of 8 is not shown in its proper scale for lack of space. The gray-colored bins are the first-ranked bins ('digit 1') within each cycle.

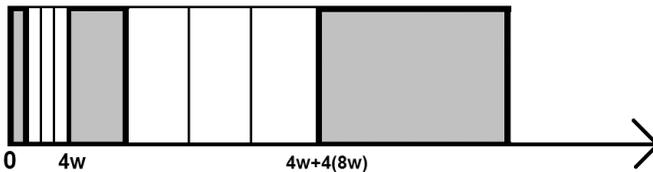
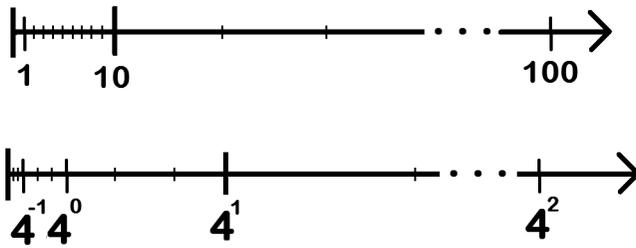


Figure 7.11 Four-Bin System Expanding by a Factor of 8 (Scheme A)



**Figure 7.12** Positional Number Systems Thought of as Particular Bin Schemes

Figure 7.12 helps in demonstrating the similarities and differences between partitioning of the x-axis by way of bin systems and by way of positional number systems along digital lines. Positional number systems contain infinitely many digital cycles which are not ‘directly’ or ‘intentionally’ placed from the 0 origin as can be done actively in bin systems. Rather, positional number systems can be thought of as creating digital cycles from the base onwards, as well as from the base backwards towards the origin in an infinite process, which in the limit ‘starts’ from the origin for all practical matter.

Upon learning of the statistician’s course of action and his intent to publish such scandalous material in a popular journal written in the local Tau Cetian language, the grand inquisitor, head of the Peace, Harmony and Suppression of Numbers Department, lodged a formal subpoena for the statistician to appear for a trial. The chief planetary statistician bore himself with grace throughout the trial. To the accusations of the inquisitor that the whole scheme smacks too much of numbers and digits, the statistician pleaded not guilty and replied that even though the scheme may resemble the system used on Earth, it has nothing to do with a number system. He calmly reminded the inquisitor that any positional number system needs expanding exactly by the value of the base, which in terms of bin schemes is expressed as  $F = D + 1$ , while he has intentionally and deliberately avoided such a socially deplorable act altogether in both his schemes. This argument was presented to the court by showing that any attempt to establish a number system based on Scheme A — thought of supposedly as base 5 number system — would fail as it expands by a factor of 8 instead of the usual 5. The same argument was given for Scheme B if thought of in terms of base 8 number system, which should not be expanding by 3 but rather by the usual 8. Another crucial argument was made reminding the court that a positional number system does not strictly start with a finite segment from the origin as in the Schemes A and B. Agreeing to

a forced confession that he dealt with forbidden numbers and digits spared the statistician the usual death penalty applied for such serious offences, and he was instead condemned to house arrest for the rest of his life. Legend has it that as he walked out of the courthouse disillusioned and bitter but still highly determined; he whispered quietly to himself and his close associates “and in spite of it all, the scheme still measures quantitatively, not numerically!”

When the news about Tau-Ceti-f formulations of these two new quantitative laws reached Earth, local data and abstract distributions were immediately checked and compared with Schemes A and B above. The results strongly confirmed what was discovered on Tau-Ceti-f, with the exceptions of U.S. county area data, the Normal, and the Uniform, which are not logarithmic digit-wise to begin with. This shocked and embarrassed Earth statisticians who always considered their number system superior and looked down upon Tau Cetians for their total lack of numbers and digits and their general backwardness. Earthlings felt deeply offended by the fact that their number system (their pride) was not even needed in order to express the universal pattern of relative quantities. It was only sheer academic honesty that forced them to accept what was clearly and consistently confirmed empirically. As events proved, the very lack of a number system actually stimulated creativity on Tau-Ceti-f and forced them to invent more general principles. The tables in Figs. 7.13 and 7.14 show the strong compliance of Earth data and distributions with the two new laws proclaimed on Tau-Ceti-f for the benefit of the whole of the galaxy.

Notes: (1) Time between earthquakes pertains to Case Study I in Chapter 11 for data on 19,452 global earthquake occurrences during 2012. (2) U.S. Population centers data pertains to Case Study II in Chapter 13 about population count of

Data Set	Bin A	Bin B	Bin C	Bin D
Time Between Earthquakes	48.3%	25.0%	15.3%	11.5%
USA Population Centers	48.9%	23.1%	16.0%	12.0%
LOG Symmetrical Triangular (1, 3, 5)	48.8%	24.1%	15.4%	11.7%
k/x over (1, 1000000)	49.3%	21.7%	16.3%	12.7%
Exponential Growth, B=1.5, F=1.01	47.9%	23.7%	15.6%	12.8%
Lognormal, Location=5, Shape=1	49.1%	23.3%	15.6%	12.1%
Lognormal, Location=9.3, Shape=1.7	48.6%	23.7%	15.8%	11.9%
Varied Data - Hill's Model	46.3%	25.3%	16.2%	12.2%
Chain U(U(U(U(0, 5666))))	47.8%	24.1%	16.1%	12.0%

Figure 7.13 Earth Data Proportions — Expanding Four-Bin System F = 8 (Scheme A)

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F	Bin G
Time Between Earthquakes	22.4%	18.1%	15.5%	13.1%	11.7%	10.1%	9.0%
USA Population Centers	22.6%	18.9%	15.4%	13.1%	10.9%	9.9%	9.1%
LOG Symmetrical Triangular (1, 3, 5)	23.0%	17.8%	15.0%	13.0%	11.5%	10.2%	9.4%
k/x over (1, 1000000)	21.5%	17.9%	15.2%	13.4%	12.5%	10.3%	9.2%
Exponential Growth, B=1.5, F=1.01	22.6%	18.0%	15.0%	12.9%	11.7%	10.4%	9.4%
Lognormal, Location=5, Shape=1	23.1%	18.1%	14.9%	13.2%	11.4%	10.2%	9.1%
Lognormal, Location=9.3, Shape=1.7	22.8%	18.3%	15.2%	12.9%	11.5%	10.2%	9.2%
Varied Data - Hill's Model	22.0%	20.1%	15.6%	13.1%	10.3%	9.5%	9.4%
Chain U(U(U(U(U(0, 5666))))))	23.2%	17.8%	15.6%	13.3%	10.8%	10.2%	9.1%

Figure 7.14 Earth Data Proportions — Expanding Seven-Bin System F = 3 (Scheme B)

19,509 cities and towns in the USA in 2009. (3) Log symmetrical triangular refers to a distribution for LOG values shaped as a triangle, starts from 1, centers on 3, ends on 5, and it gives rise to data by simply calculating  $10^{\text{Triangular}}$ . (4) Only the first 10,000 elements of the exponential 1% growth from a base of 1.5 are considered. (5) Varied Data Hill's Model refers to the 34,269 values randomly obtained from 70 different Earthly sources and topics found on their electronic web system mentioned in Chapter 110 of Section 6, Case Study XI. (6) The chain of distributions refers to 20,000 realizations from computer simulations of the scheme: Uniform(0, Uniform(0, Uniform(0, Uniform(0, Uniform(0, 5666))))).

The tables in Figs. 7.13 and 7.14 clearly tell the story of a very consistent fall in the densities of all of these logarithmic data sets and distributions (in the aggregate over the entire range), no matter how many bins are set to measure it. Crucially, for a given set of D, F, S, W values, that rate of fall (i.e. bin proportions vector) is remarkably uniform across all logarithmic data types and distributions, with minor variations.

## NON-EXPANDING BIN SYSTEM MEASURING FALL IN K/X DISTRIBUTION

---

Postulating that the generic pattern in how relative quantities are found in nature is such that the frequency of quantitative occurrences is inversely proportional to quantity, we are then led to explore results from bin systems fitting  $k/x$  distribution, the density curve whose arrangement of relative quantities constitutes exactly such a pattern. Three cases shall be examined: non-expanding, once-expanding, and twice-expanding bin systems. The goal is to arrive at a general expression for bin proportions in the  $k/x$  case which would then be checked against empirical bin results of real data, assuming that a consistent and steady law could be mathematically found for the  $k/x$  case.

The chart in Fig. 7.15 depicts the generic Benford density curve of  $k/x$  distribution defined over  $(w, (D+1)w)$ . This represents a basic or unitary bin system with no expansion, having only one cycle. Six features are involved in this construction:

- (I) Avoidance of an upward explosion start of the  $k/x$  density at the origin 0 which would have been undefined due to a division by 0.
- (II) Equal spacing (width) of all bins.
- (III) Equality between bin width and the separation of the defined range from the 0 origin. Namely, that the length of the step from the origin to the launch of  $k/x$  is also the width of each bin. Algebraically,  $(2w - w) = (w - 0)$ .
- (IV) No coordination is employed or attempted whatsoever with any number system or digits on the x-axis below.
- (V) No assumption or constraint is made about the value of width  $w$ , which is left completely flexible to take on any value whatsoever, including fractional values.
- (VI) Only positive numbers are involved.

Equating the entire area to one, we obtained  $\int k/x \, dx = 1$  over  $[w, (D+1)w]$ , therefore  $k[\ln((D+1)w) - \ln(w)] = 1$ , or  $k[\ln(D+1) + \ln(w) - \ln(w)] = 1$ , so that

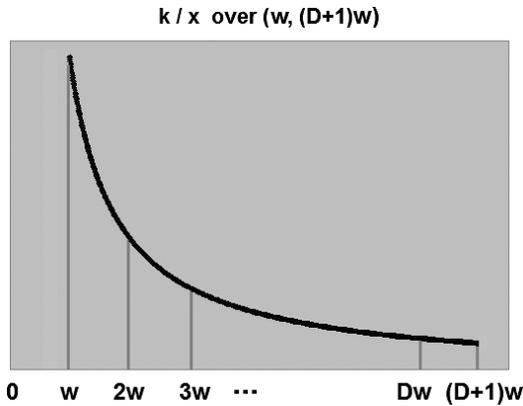


Figure 7.15 Fall in  $k/x$  Distribution as Measured by a Singular Bin System Cycle

$k[\ln(D+1)] = 1$ , and  $k = 1/\ln(D+1)$ . Evaluating the portion of area hanging over bin #d (with d running from 1 to D, as in digits), we obtain  $P(d) = \int k/x \, dx$  over  $[d*w, (d+1)*w]$ , namely  $P(d) = [1/\ln(D+1)] * [\ln(d+1) + \ln(w) - \ln(d) - \ln(w)] = [1/\ln(D+1)] * [\ln(d+1) - \ln(d)]$ , or  $P(d) = [1/\ln(D+1)] * \ln[(d+1)/(d)]$ , and finally  **$P(d) = \ln(1+1/d)/\ln(D+1)$** .

Double application of the logarithmic identity  $\text{LOG}_A X = \text{LOG}_B X / \text{LOG}_B A$  yields  $[\log(1+1/d)/\log(e)] / [\log(D+1)/\log(e)]$  where log is of the decimal base 10, and finally  **$P(d) = \log(1+1/d)/\log(D+1)$** . Hence, the above bin-proportion expression is perfectly compatible with the more general expression of Benford’s Law encompassing other bases since D+1 is conceptually the equivalent value of the base in any number system. Since D represents the number of bins, it then also represents the number of first digits when the bin scheme is viewed in the context of a number system. Hence, D+1 expresses the value of the base in such a vista. Yet, this particularly short scheme for  $k/x$  distribution must be considered with severe reservation due to its intrinsic limitation — lacking expansions of the cycles. Without expansions to ever larger bin cycles, consistency of results is in doubt.

Nonetheless, the remarkable philosophical and conceptual implication of this result is that the form of Benford’s Law in the expression  $\log(1+1/d)/\log(\text{base})$  serves also (or mainly) as a general quantitative proportional law outside any digital framework in the restricted case of  $k/x$  without any bin expansion. This is so since width  $w$  cancels out and drops from the calculations. As a result,  $w$  could take

on any value without affecting the above expression. Crucially, an odd mixture of distinct leading digits reside harmoniously within each bin — depending on the value of  $w$  chosen — yet bin proportions are always as in  $\log(1+1/d)/\log(D+1)$ ! The fact that bin results here are independent of the width  $w$  is significant also in another way, since this lends the bin scheme for  $k/x$  universality and consistency. Equivalently stated,  $\log(1+1/d)/\log(\text{base})$  serves also to express relative quantities and the rate of fall in the density of  $k/x$  defined over some restricted range.

It should be noted that the above setup of  $k/x$  distribution defined over the interval  $[w, (D+1)w]$  is considered perfectly logarithmic in digital Benford’s Law sense for any base  $B$ , where  $D$  signifies the number of all first-digit possibilities (such as nine first digits in base 10 number system). This is so because exponent difference of this range is  $\text{LOG}_B((D+1)w) - \text{LOG}_B(w) = \text{LOG}_B(D+1)$ , and since  $D+1$  represents the base as well, this reduces to  $\text{LOG}_B(B)$ , namely 1, which is an integral number, and therefore the distribution is logarithmic.

As it happened, another Clarikian physical data set corresponding perfectly to  $k/x$  defined over  $(20, 200)$  is also clearly observable on both Tau-Ceti-f and Earth. On Tau-Ceti-f this Clarikian data set was analyzed as a non-expanding bin scheme with  $w = 20$  and  $D = 9$ . The nine bins measuring the fall are  $\{(20, 40), (40, 60), (60, 80), (80, 100), (100, 120), (120, 140), (140, 160), (160, 180), (180, 200)\}$ . Figure 7.16 depicts the scheme performed on Tau-Ceti-f. It was discovered that 30.1% of data falls within the first bin  $(20, 40)$ , 17.6% of data falls within  $(40, 60)$ , only 4.6% falls within the last bin  $(180, 200)$ , and that all proportions somehow

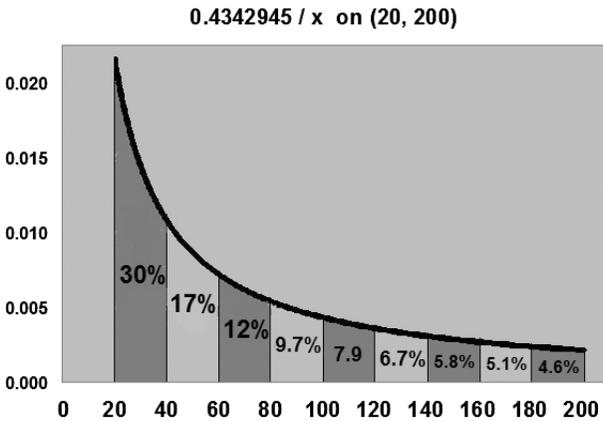


Figure 7.16 Bins Measuring Relative Quantities and Fall in  $k/x$  (20, 200) — Tau-Ceti-f

perfectly fit the expression  $\text{LOG}_{10}(1 + 1/d)$  as in that old and forbidden Zikuma's Law, except that base 10 is applied, not base 9. Width equality between bins is a basic feature in all bin schemes, and thus the first bin (20, 40) is as wide as the last one (180, 200).

On Earth, this Clarikian physical data set was instinctively analyzed in terms of the well-known law of Benford, satisfying their supreme obsession with digits. Data was found to perfectly comply with the law. Figure 7.17 depicts the very different view from Earth about the same physical data set, where the focus was to calculate areas according to first significant digits. Earth was not troubled at all that the sizes of the various sub-sections on the x-axis were not of the same width, nor was there any objection to placing the 'first' of the first leading digit (i.e. 1) way in the back being the last x-axis section within the whole partition. Both planets have communicated their own perspectives about this particular Clarikian data set to the chief galactic statistician, formally requesting an approval. This caused a great deal of friction between the two planets and gave the splendidly clothed generals another grand casus belli against Earth. It took the very firm reprimand and severe dressing down of all of them (literally) by the dictator to save the galactic peace. This author is siding wholeheartedly with Tau-Ceti-f on this issue as a matter of academic principle and firmly rejects Earth's numerocentricity, and in spite of his Earthly origin, not only due to enthusiastic preference of quantities over digits, but also because of the simplicity, unity, evenness, and order found in Tau-Ceti-f approach. Quick simultaneous glances at Figs. 7.16 and 7.17 for

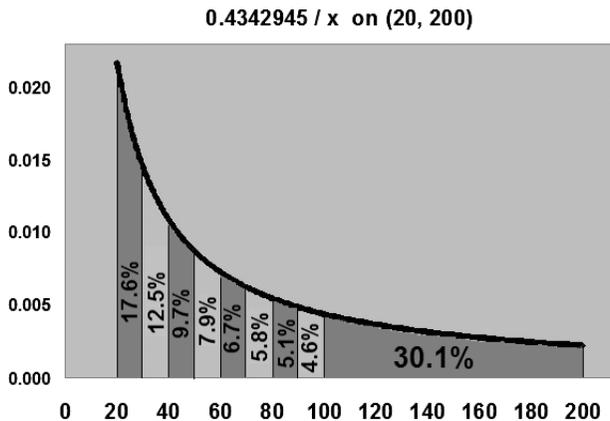


Figure 7.17 Digits and Benford's Law Calculated for  $k/x$  on (20, 200) — Earth

comparison would convince anyone (regardless of planetary ancestry) of the relative attractiveness of 7.16.

When the constraint of equality between the width of the bins and the separation from the origin is violated, results differ considerably. For example, for  $k/x$  defined over (1, 181), the separation is merely one unit, and a non-expanding nine-bin system where the width of each bin is 20 units, divides the range into (1, 21), (21, 41), (41, 61), (61, 81), (81, 101), (101, 121), (121, 141), (141, 161), (161, 181). Here, bin proportions are highly skewed in favour of 'low bins' at {58.6%, 12.9%, 7.6%, 5.5%, 4.2%, 3.5%, 2.9%, 2.6%, 2.3%}.

For  $k/x$  defined over (420, 600), the large separation is 420 units, and a non-expanding nine-bin system where the width of each bin is 20 units, divides the range into (420, 440), (440, 460), (460, 480), (480, 500), ... , (580, 600). Here, bin proportions are nearly even at {13.0%, 12.5%, 11.9%, 11.4%, 11.0%, 10.6%, 10.2%, 9.8%, 9.5%}.

Intuitively, when an imaginary fixed 180-unit segment of  $k/x$  defined over ( $\approx 0$ ,  $\approx \infty$ ) is focused on, it appears sloping sharply near the origin (thus bin proportions there are highly skewed), less so a bit further to the right, and it appears nearly flat much further on the far right (thus bin proportions there are nearly even).

**ONCE-EXPANDING BIN SYSTEM MEASURING FALL IN K/X DISTRIBUTION**

Figure 7.18 depicts the generic Benford density curve of  $k/x$  distribution defined over  $(w, (D+1)w + DFw)$ . This represents a bin system with a single expansion where original bin's width  $w$  is being inflated by  $F$  on the second cycle. That is, where the width of each bin in the second cycle is  $Fw$  instead of  $w$ .

Equating the entire area to one, we obtained:

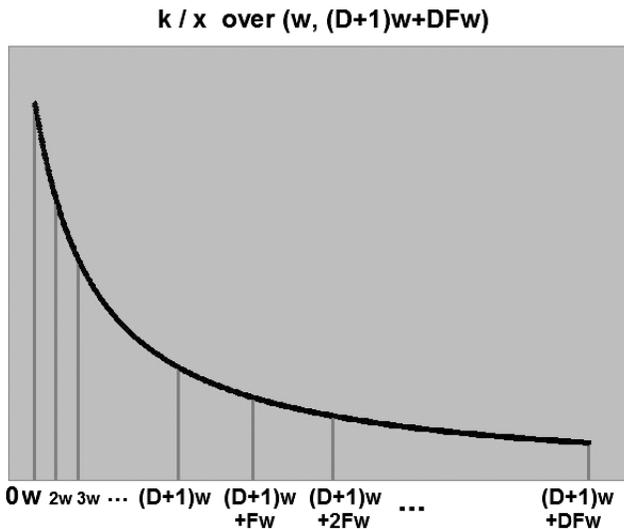
$$\int k/x \, dx = 1 \text{ over } [w, (D+1)w + DFw]$$

$$k[\ln(w*(D+1) + (DF)) - \ln(w)] = 1$$

$$k[\ln(w) + \ln[(D+1) + (DF)] - \ln(w)] = 1$$

$$k[\ln[(D+1) + (DF)]] = 1$$

$$k = 1/\ln(1 + D + DF).$$



**Figure 7.18** Fall in  $k/x$  Distribution as Measured by Once-Expanding Bin System

Evaluating the **first** portion of area (first cycle) hanging over bin #d (d running from 1 to D as in digits), we obtain:

$$P1(d) = \int k/x \, dx \text{ over } [d*w, (d+1)*w]$$

$$P1(d) = [1/\ln(1 + D + DF)] * [\ln(d+1) + \ln(w) - \ln(d) - \ln(w)]$$

$$P1(d) = [1/\ln(1 + D + DF)] * [\ln(d+1) - \ln(d)].$$

Evaluating the **second** portion of area (second cycle) hanging over bin #d (with d running from 1 to D, as in digits), we obtain:

$$P2(d) = \int k/x \, dx \text{ over } [(D+1)w + (d-1)*Fw, (D+1)w + (d)*Fw]$$

$$P2(d) = [1/\ln(1 + D + DF)] * [\ln((D+1) + dF) + \ln(w) - \ln((D+1) + (d-1)F) - \ln(w)]$$

$$P2(d) = [1/\ln(1 + D + DF)] * [\ln((D+1) + dF) - \ln((D+1) + (d-1)F)]$$

Combining both areas, namely  $P(d) = P1(d) + P2(d)$ , we get:

$$P(d) = [1/\ln(1 + D + DF)] * [\ln(d+1) - \ln(d) + \ln((D+1) + dF) - \ln((D+1) + (d-1)F)]$$

Applying the identity  $\text{LOG}(A) - \text{LOG}(B) = \text{LOG}(A/B)$  we finally get:

$$P(d) = [\ln(1 + 1/d) + \ln[(1+D+dF)/(1+D+(d-1)F)]] / [\ln(1 + D + DF)]$$

## ONCE-EXPANDING BINS FOR $k/x$ REDUCES TO BENFORD WHEN $F = D + 1$

---

The bin system set up for  $k/x$  distribution with one expansion (two cycles) reduces to Benford's Law whenever inflation factor  $F$  is made equal to  $(D + 1)$  as in all proper number systems. This implies that whenever  $F = D + 1$ , once-expanding bin system reduces to non-expanding bin system. Another way of viewing the implication here is to note that a singular expansion — the doubling of the number of bins — does not change bin proportions in any way for  $k/x$  distribution whenever  $F = D + 1$ .

To prove the assertion,  $(D + 1)$  is simply substituted for  $F$  in the final  $P(d)$  expression of the previous chapter, hence:

$$P(d) = [\ln(1 + 1/d) + \ln[(1+D+dF)/(1+D+(d-1)F)]] / [\ln(1 + D + DF)]$$

$$P(d) = [\ln(1 + 1/d) + \ln[(1+D+d(D + 1))/(1+D+(d-1)(D + 1))]] / [\ln(1 + D + D(D + 1))]$$

$$P(d) = [\ln(1 + 1/d) + \ln[((1+D)(1 + d))/((1+D)(1+(d-1)))] / [\ln(1 + D + D^2 + D)]$$

$$P(d) = [\ln(1 + 1/d) + \ln[((1+D)(1 + d))/((1+D)(d))]] / [\ln(1 + 2D + D^2)]$$

$$P(d) = [\ln(1 + 1/d) + \ln[(1 + d)/(d)]] / [\ln((D + 1)^2)]$$

$$P(d) = [\ln(1 + 1/d) + \ln[(1 + 1/d)]] / [\ln((D + 1)^2)]$$

$$P(d) = 2 * \ln(1 + 1/d) / \ln((D + 1)^2)$$

$$P(d) = 2 * \ln(1 + 1/d) / 2 * \ln(D + 1)$$

$$P(d) = \ln(1 + 1/d) / \ln(D + 1)$$

$$P(d) = \log(1 + 1/d) / \log(D + 1)$$

$$P(d) = \log(1 + 1/d) / \log(\text{Base})$$

$$P(d) = \text{Benford's Law}$$

## TWICE-EXPANDING BIN SYSTEM MEASURING FALL IN K/X DISTRIBUTION

The chart in Fig. 7.19 depicts the generic Benford density curve of  $k/x$  distribution defined over  $(w, (D+1)w + DFw + DF^2w)$ . This represents a bin system with two expansions, where original bin width  $w$  is being inflated by  $F$  on the second cycle, and inflated by  $F^2$  on the third cycle (or equivalently, inflated by  $F$  again on the third cycle from the base width of  $Fw$  of the preceding second cycle).

Equating the entire area to one, we obtained  $\int k/x \, dx = 1$  over the range of  $[w, (D+1)w + DFw + DF^2w]$ , hence  $k[\ln(w) + \ln[(D+1) + DF + DF^2] - \ln(w)] = 1$ , or  $k[\ln[(D+1) + DF + DF^2]] = 1$ , and finally  $k = 1 / \ln(1 + D + DF + DF^2)$ .

Evaluating the **first and second** portion of areas yields the same results as in the once-expanding bin system except for the different  $k$  constant expression

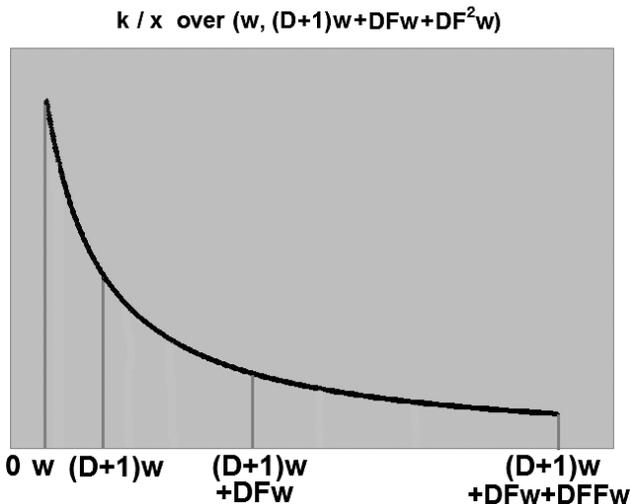


Figure 7.19 Fall in  $k/x$  Distribution as Measured by Twice-Expanding Bin System

here. Evaluating the **third** portion of area hanging over bin #d (with d running from 1 to D, as in digits), we obtain:

$$P3(d) = \int k/x \, dx \text{ over } [(D+1)w + DFw + (d - 1)*F^2w, (D+1)w + DFw + (d)*F^2w]$$

$$P3(d) = k (\ln[w] + \ln[(D+1) + DF + (d)*F^2] - \ln[w] - \ln[(D+1) + DF + (d - 1)*F^2])$$

$$P3(d) = k (\ln[(D+1) + DF + (d)*F^2] - \ln[(D+1) + DF + (d - 1)*F^2])$$

$$P3(d) = k * \ln([(D+1) + DF + (d)*F^2]/[(D+1) + DF + (d - 1)*F^2])$$

Combining all three areas, namely  $P(d) = P1(d) + P2(d) + P3(d)$ , we finally get:

$$P(d) = k*\ln(1 + 1/d) + k*\ln([1+D+dF]/[1+D+(d - 1)F]) +$$

$$k*\ln([(D+1) + DF + (d)*F^2]/[(D+1) + DF + (d - 1)*F^2])$$

$$P(d) = \frac{\ln(1 + 1/d) + \ln\left(\frac{[1 + D + (d)F]}{[1 + D + (d - 1)F]}\right) + \ln\left(\frac{[1 + D + DF + (d)*F^2]}{[1 + D + DF + (d - 1)*F^2]}\right)}{\ln(1 + D + DF + DF^2)}$$

## TWICE-EXPANDING BINS FOR K/X REDUCES TO BENFORD WHEN $F = D + 1$

Twice-expanding bin system setup for  $k/x$  distribution also reduces to Benford's Law whenever inflation factor  $F$  is made equal to  $(D + 1)$  as in all proper number systems. In other words, whenever  $F = D + 1$ , twice-expanding bin system reduces to non-expanding bin system. The implication here is that the very act of doubly expanding the bins (i.e. the tripling of the number of bins) does not change bin proportions whenever  $F = D + 1$ .

To prove the assertion,  $(D + 1)$  is simply substituted for  $F$  in the final  $P(d)$  expression of the previous chapter, hence:

$$\frac{\ln(1 + 1/d) + \ln\left(\frac{[1 + D + (d)(D+1)]}{[1 + D + (d-1)(D+1)]}\right) + \ln\left(\frac{[1 + D + D(D+1) + (d)*(D+1)^2]}{[1 + D + D(D+1) + (d-1)*(D+1)^2]}\right)}{\ln(1 + D + D(D+1) + D(D+1)^2)}$$

$$\frac{\ln(1 + 1/d) + \ln\left(\frac{[(d+1)(D+1)]}{[(d-1+1)(D+1)]}\right) + \ln\left(\frac{[(1+D)(D+1) + (d)*(D+1)^2]}{[(1+D)(D+1) + (d-1)*(D+1)^2]}\right)}{\ln((1 + D)(D+1) + D(D+1)^2)}$$

$$\frac{\ln(1 + 1/d) + \ln\left(\frac{[(d+1)(D+1)]}{[(d)(D+1)]}\right) + \ln\left(\frac{[(1+D)*[(D+1) + (d)*(D+1)]]}{[(1+D)*[(D+1) + (d-1)*(D+1)]]}\right)}{\ln((1+D)*[(D+1) + D(D+1)])}$$

$$\frac{\ln(1 + 1/d) + \ln\left(\frac{[d+1]}{[d]}\right) + \ln\left(\frac{[(D+1) + (d)*(D+1)]}{[(D+1) + (d-1)*(D+1)]}\right)}{\ln((1+D)*[(D+1)(1+D)])}$$

$$\frac{\ln\left(\frac{[d+1]}{[d]}\right) + \ln\left(\frac{[d+1]}{[d]}\right) + \ln\left(\frac{[(D+1)[1+d]]}{[(D+1)[1+(d-1)]}\right)}{\ln((1+D)^3)} = \frac{3*\ln\left(\frac{[1+d]}{[d]}\right)}{3*\ln(1+D)} = \mathbf{BL}$$

Hence, no contradictions or paradoxes arise in the construction of the bin system, at least not in the context of compatibility with our number system and its associated Benford’s Law. Another thorough check on triply-expanding bin system (four bin cycles) for  $k/x$  distribution was performed (details not shown here), confirming the same reduction to Benford’s Law in the case of  $F = D + 1$ , as shown here in the previous two cases. While this surely does not constitute a formal proof, it strongly suggests that any  $N$ -expanding (and even infinitely-expanding) bin system for  $k/x$  would yield the same reduction to the law given that  $F = D + 1$ .

## INFINITELY EXPANDING BIN SYSTEM MEASURING FALL IN K/X

---

Algebraic expressions for bin proportions of  $k/x$  distribution for higher expansion orders perfectly follow the above (clear) pattern as a sequence of ever increasing terms in the numerator and in the denominator. The first four elements of this infinite sequence, beginning with a non-expanding bin system and ending with a bin system having three expansions, are as follows:

$$\frac{\ln\left(\frac{[1 + (d)]}{[1 + (d-1)]}\right)}{\ln(1 + D)}$$

$$\frac{\ln\left(\frac{[1 + (d)]}{[1 + (d-1)]}\right) + \ln\left(\frac{[1 + D + (d)F]}{[1 + D + (d-1)F]}\right)}{\ln(1 + D + DF)}$$

$$\frac{\ln\left(\frac{[1 + (d)]}{[1 + (d-1)]}\right) + \ln\left(\frac{[1 + D + (d)F]}{[1 + D + (d-1)F]}\right) + \ln\left(\frac{[1 + D + DF + (d)F^2]}{[1 + D + DF + (d-1)F^2]}\right)}{\ln(1 + D + DF + DF^2)}$$

$$\frac{\ln\left(\frac{[1+(d)]}{[1+(d-1)]}\right) + \ln\left(\frac{[1+D+(d)F]}{[1+D+(d-1)F]}\right) + \ln\left(\frac{[1+D+DF+(d)F^2]}{[1+D+DF+(d-1)F^2]}\right) + \ln\left(\frac{[1+D+DF+DF^2+(d)F^3]}{[1+D+DF+DF^2+(d-1)F^3]}\right)}{\ln(1 + D + DF + DF^2 + DF^3)}$$

The assumption that this infinite sequence of bin proportions of  $k/x$  distribution over an infinite range on the  $x$ -axis converges has been strongly suggested by

way of (finite) computer simulations for a large variety of F and D values, lending support to the extrapolation that convergence occurs for all possible combinations of F and D values. Algebraic manipulations performed in the chapter after the next one yield a closed-form expression for the limit of this infinite algebraic sequence in terms of F and D, formally guaranteeing convergence.

## CONFIRMATION MATCHING K/X FALL WITH EMPIRICAL BINS ON REAL DATA

---

The moment of truth has arrived: real-life data viewed through the prism of bin schemes must be checked against the converging limit of the algebraic sequence derived earlier regarding the fall in the generic density  $k/x$  viewed through an infinitely expanded bin scheme. Needless to say, only corresponding bin systems having the same values of  $F$  and  $D$  should be used for valid comparisons as empirical confirmation.

Let us compare the two empirical bin schemes of real Earth data and abstract distributions as shown in Figs. 7.13 and 7.14 with equivalent infinitely expanded bins superimposed on the  $k/x$  curve. The limit of the generic  $k/x$  is calculated with the aid of a computer program using 400 bin cycles, although quick convergence is strongly suggested by the machine much earlier after the first 100 cycles or so. There exists still no formal mathematical assurance of the sequence's convergence at this stage in the development of our bin theory, yet it shall be taken for granted, and this assumption shall be vindicated in the next chapter where it is proven. The two averages of the bin results from all nine data sets and distributions, and their equivalent  $k/x$  bin computerized results are as follows:

Scheme A:

$$D = 4 \quad F = 8 \quad S = 0 \quad W = 0.0008$$

Average proportions of nine sets = {48.3%, 23.8%, 15.9%, 12.0%}

Limit of infinite sequence of  $k/x$  = {48.6%, 23.7%, 15.8%, 11.9%}

Scheme B:

$$D = 7 \quad F = 3 \quad S = 0 \quad W = 0.0008$$

Avg of nine sets = {22.6%, 18.4%, 15.2%, 13.1%, 11.3%, 10.1%, 9.3%}

Limit of  $k/x$  seq. = {22.9%, 18.3%, 15.2%, 13.0%, 11.4%, 10.1%, 9.1%}

Hooray, it is confirmed! Moreover, deviation from the average of the nine data sets within each scheme is extremely low, as seen in Figs. 7.13 and 7.14.

**The limit of the sequence of algebraic expressions for the infinitely expanded  $k/x$  distribution should now be construed as constituting the general law governing occurrences of relative quantities in all logarithmic data sets** (as a function of the manner in which such a concept is measured, namely as a function of  $D$  and  $F$ , with  $S$  and  $w$  being insignificant so long as they are confined to very small values).

An important aspect in all expanding bin systems for  $k/x$  distribution is independence of results on width  $w$  as it drops out in the calculations and thus could take on any value without affecting resultant expression whatsoever. Yet caution should be exercised when real-life data is concerned and where a large value of  $w$  totally distorts results. The dichotomy emanates from the fact that by definition, for all bin schemes on  $k/x$  there exists no values/data between 0 and  $w$ , while for real-life data a huge portion of data might be hiding there. In other words,  $k/x$  schemes start at  $w$  and terminate at infinity while real data may start at or near 0 and always terminates at a finite maximum value. Therefore, setting  $w$  equal to some very small fractional value for real data is advisable if compatibility and correspondence between the generic  $k/x$  distribution and real life data set is desired, as in the theoretical framework developed here.

## CLOSED FORM EXPRESSION FOR THE LIMIT OF THE INFINITE SEQUENCE

---

With crucial assistance from the distinguished mathematician George Andrews a closed form (analytical) expression for the limit of the infinite sequence of k/x bin scheme model is obtained in the  $F > 1$  case, enabling us to succinctly express the general law of relative quantities. The assumption  $F \neq 1$  is solely an initial restriction enabling progress in the reduction and it shall be relaxed later.

The fourth term of the sequence expressed earlier, denoted as  $S_4$ , is:

$$\frac{\ln\left(\frac{[1+(d)]}{[1+(d-1)]}\right) + \ln\left(\frac{[1+D+(d)F]}{[1+D+(d-1)F]}\right) + \ln\left(\frac{[1+D+DF+(d)F^2]}{[1+D+DF+(d-1)F^2]}\right) + \ln\left(\frac{[1+D+DF+DF^2+(d)F^3]}{[1+D+DF+DF^2+(d-1)F^3]}\right)}{\ln(1 + D + DF + DF^2 + DF^3)}$$

Employing the **finite** geometric formula for the terms involving F, namely:

$$1 + X + X^2 + X^3 + \dots + X^N = (X^{N+1} - 1)/(X - 1) ,$$

the nth term in the sequence is then:

$$S_N = \frac{\sum_{j=0}^{j=N-1} \ln \left( \frac{1 + \frac{D(F^j - 1)}{(F - 1)} + (d)F^j}{1 + \frac{D(F^j - 1)}{(F - 1)} + (d-1)F^j} \right)}{\ln \left( 1 + \frac{D(F^N - 1)}{(F - 1)} \right)}$$

Pulling together all the coefficients of  $F^{\text{POWER}}$ , we get:

$$S_N = \frac{\sum_{j=0}^{j=N-1} \ln \left( \frac{1 + \left( \frac{D + (d)(F-1)}{(F-1)} \right) F^j - \frac{D}{F-1}}{1 + \left( \frac{D + (d-1)(F-1)}{(F-1)} \right) F^j - \frac{D}{F-1}} \right)}{\ln \left( 1 + \left( \frac{D}{(F-1)} \right) F^N - \frac{D}{F-1} \right)}$$

In order to obtain a more compact expression, let us define:

$$A = \frac{D + d(F-1)}{F-1}$$

$$B = \frac{D + (d-1)(F-1)}{F-1}$$

$$C = \frac{D}{F-1}$$

$$E = 1 - \frac{D}{F-1}$$

$$S_N = \frac{\sum_{j=0}^{j=N-1} \ln \left( \frac{AF^j + E}{BF^j + E} \right)}{\ln(CF^N + E)}$$

Since in our context  $F \geq 1$ , namely either 1 as in flat bin schemes, or larger than 1 as in normal expanding bin schemes yielding quantitative laws, there is no hope of obtaining any obvious convergence in terms such as  $F^j$  or  $F^N$ . Hence, we define  $f = 1/F$ , creating a quantity  $f$  such that  $0 < f \leq 1$  holds and which may hopefully let terms such as  $f^j$  or  $f^N$  converge.

$$S_N = \frac{\sum_{j=0}^{j=N-1} \ln \left( \frac{A \frac{1}{f^j} + E}{B \frac{1}{f^j} + E} \right)}{\ln \left( C \frac{1}{f^N} + E \right)}$$

$$S_N = \frac{\sum_{j=0}^{j=N-1} \ln \left( \frac{A \frac{1}{f^j} + E}{B \frac{1}{f^j} + E} * \frac{f^j}{f^j} \right)}{\ln \left( C * \frac{1}{f^N} + E * \frac{C}{C} * \frac{f^N}{f^N} \right)}$$

$$S_N = \frac{\sum_{j=0}^{j=N-1} \ln \left( \frac{A + E * f^j}{B + E * f^j} \right)}{\ln \left( C * \frac{1}{f^N} * \left( 1 + E * \frac{f^N}{C} \right) \right)}$$

$$S_N = \frac{\sum_{j=0}^{j=N-1} \ln \left( \frac{\left( 1 + \frac{E}{A} * f^j \right) A}{\left( 1 + \frac{E}{B} * f^j \right) B} \right)}{\ln \left( C * F^N * \left( 1 + \frac{E}{C} f^N \right) \right)}$$

$$S_N = \frac{N * \ln \left( \frac{A}{B} \right) + \ln \left( \prod_{j=0}^{N-1} \left( \frac{1 + \frac{E}{A} * f^j}{1 + \frac{E}{B} * f^j} \right) \right)}{N * \ln(F) + \ln(C) + \ln \left( 1 + \frac{E}{C} f^N \right)}$$

It is only at this late stage that we let N go to infinity!

For  $F > 1$  in the normal case of expanding bin scheme,  $0 < f < 1$ , therefore

$$\prod_{j=0}^{\infty} \frac{\left(1 + \frac{E}{A} * f^j\right)}{\left(1 + \frac{E}{B} * f^j\right)}$$

is a convergent infinite product since  $\sum_{j=0}^{\infty} f^j$  is converging. The term  $\ln\left(1 + \frac{E}{C} f^N\right)$  is zero as  $N \rightarrow \infty$ . The term  $\ln(C)$  is  $\ln\left(\frac{D}{F-1}\right)$ , and it is also finite as  $N \rightarrow \infty$ , finally:

$$N \rightarrow \infty \quad S_N = \left(\frac{\ln\left(\frac{A}{B}\right)}{\ln(F)}\right)$$

and using the definitions of A and B above, the general relative quantities law is then

$$\frac{\ln\left(\frac{\left(\frac{D+d(F-1)}{F-1}\right)}{\left(\frac{D+(d-1)(F-1)}{F-1}\right)}\right)}{\ln(F)},$$

which is further reduced by canceling out the two  $(F - 1)$  terms in the numerator to arrive at:

**The General Law:** 
$$\frac{\ln\left(\frac{D+d(F-1)}{D+(d-1)(F-1)}\right)}{\ln(F)}$$

To verify that digital Benford’s Law is simply a special case and a consequence of the general law when bin schemes are constructed under the constraint  $F = D + 1$ , the term F is then substituted by  $D + 1$  everywhere in expression of the general law:

$$\begin{aligned} \text{GL} &= \frac{\ln\left(\frac{D+d(D+1-1)}{D+(d-1)(D+1-1)}\right)}{\ln(D+1)} = \frac{\ln\left(\frac{D+d(D)}{D+(d-1)(D)}\right)}{\ln(D+1)} = \frac{\ln\left(\frac{1+d}{1+(d-1)}\right)}{\ln(D+1)} \\ &= \frac{\ln\left(1+\frac{1}{d}\right)}{\ln(D+1)} = \frac{\ln\left(1+\frac{1}{d}\right)}{\ln(\text{BASE})} = \frac{\text{LOG}\left(1+\frac{1}{d}\right)}{\text{LOG}(\text{BASE})} = \text{BL} \end{aligned}$$

For a number system with base 10, we get:

$$\text{BL} = \frac{\text{LOG}\left(1 + \frac{1}{d}\right)}{\text{LOG}(10)} = \frac{\text{LOG}\left(1 + \frac{1}{d}\right)}{1} = \text{LOG}_{10}\left(1 + \frac{1}{d}\right)$$

## CLOSED FORM EXPRESSION FOR THE LIMIT IN THE FLAT CASE $F = 1$

---

Let us now turn our attention to the case of flat bin schemes with  $F = 1$ , and attempt to arrive at a closed form expression. The 4th term of the sequence denoted as  $S_4$  is:

$$\frac{\ln\left(\frac{[1+(d)]}{[1+(d-1)]}\right) + \ln\left(\frac{[1+D+(d)]}{[1+D+(d-1)]}\right) + \ln\left(\frac{[1+D+D+(d)]}{[1+D+D+(d-1)]}\right) + \ln\left(\frac{[1+D+D+D+(d)]}{[1+D+D+D+(d-1)]}\right)}{\ln(1 + D + D + D + D)}$$

$$S_N = \frac{\sum_{j=0}^{j=N-1} \ln\left(\frac{1 + D * j + (d)}{1 + D * j + (d-1)}\right)}{\ln(1 + N * D)}$$

$$S_N = \frac{\sum_{j=0}^{j=N-1} \ln\left(\frac{1 + D * j + d}{D * j + d}\right)}{\ln\left(N * \left(\frac{1}{N} + D\right)\right)}$$

As  $N \rightarrow \infty$ , the numerator (NU) can be roughly evaluated by using the Integral Test Approximation.

$$\begin{aligned} \text{NU} &= \int_0^N \ln\left(\frac{1 + Dx + d}{Dx + d}\right) dx \\ &= \int_0^N [\ln(1 + Dx + d) - \ln(Dx + d)] dx \\ &= \int_0^N \ln(1 + Dx + d) dx - \int_0^N \ln(Dx + d) dx \end{aligned}$$

Let  $u = (1 + Dx + d)$ , implying that  $du/dx = D$

Let  $z = (Dx + d)$ , implying that  $dz/dx = D$

$$NU = \int_{1+d}^{1+DN+d} \ln(u) \frac{1}{D} du - \int_d^{ND+d} \ln(z) \frac{1}{D} dz$$

Using  $\int \ln(x) dx = x * \ln(x) - x + C$ , we evaluate the numerator as :

$$\begin{aligned} & + \frac{1}{D}(1 + DN + d)\ln(1 + DN + d) - \frac{1}{D}(1 + DN + d) \\ & - \frac{1}{D}(1 + d)\ln(1 + d) + \frac{1}{D}(1 + d) \\ & - \frac{1}{D}(DN + d)\ln(DN + d) + \frac{1}{D}(DN + d) \\ & + \frac{1}{D}(d)\ln(d) - \frac{1}{D}(d) \end{aligned}$$

The terms not involving N at all (in gray color) are negligible in the limit as N goes to infinity and can be omitted, hence we get:

$$\begin{aligned} & + \frac{1}{D}(1 + DN + d)\ln(1 + DN + d) - \frac{1}{D}(DN) \\ & - \frac{1}{D}(DN + d)\ln(DN + d) + \frac{1}{D}(DN) \\ \hline & + \frac{1}{D}(1 + DN + d)\ln(1 + DN + d) - \frac{1}{D}(DN + d)\ln(DN + d) \\ \hline & \frac{1}{D} * [(1)\ln(1 + DN + d) + (DN + d)\ln(1 + DN + d) - (DN + d)\ln(DN + d)] \\ \hline & \frac{1}{D} * \left[ \ln(DN) + (DN + d) * \ln\left(\frac{(1 + DN + d)}{(DN + d)}\right) \right] \\ \hline & \frac{1}{D} * \left[ \ln(D) + \ln(N) + (DN + d) * \ln\left(1 + \frac{1}{(DN + d)}\right) \right] \\ \hline & NU = \frac{1}{D} * \left[ \ln(N) + \ln\left(\left(1 + \frac{1}{(DN + d)}\right)^{(DN + d)}\right) \right] \\ \hline \end{aligned}$$

The inner expression inside the natural logarithm on the right is simply Euler's number **e** (the exponential constant) defined as  $\lim_{n \rightarrow \infty} (1 + \frac{1}{n})^n$  which converges

to the finite 2.71828 value. Hence, the whole ratio of the numerator divided by the denominator in the limit as  $N$  approaches infinity is:

$$S_{N \rightarrow \infty} = \frac{\frac{1}{D} [\ln(N) + \ln(e)]}{\ln(N) + \ln\left(\frac{1}{N} + D\right)}$$

$$S_{N \rightarrow \infty} = \frac{\frac{1}{D} \ln(N)}{\ln(N)} = \frac{1}{D}$$

Namely, bin equality for flat bin schemes with  $F = 1$ , regardless what value  $D$  takes. George Andrews suggests an alternative and concise proof in the  $F = 1$  case, derived directly from the general result of the proof in the  $F > 1$  case via L'Hopital's Rule, in spite of the fact that  $F \neq 1$  was a necessary assumption in parts of the derivation there [such as A, B, C, and E being undefined when  $F$  is 1, and violating the geometric formula restriction on common ratio not assuming the value of 1]. Andrews' approach is to let  $N$  approach infinity first as in the general  $F \neq 1$  proof, and only then let  $F$  approach 1 in the expression of the General Law. Direct substitution by 1 for  $F$  in the expression of the General Law yields the indeterminate form:

$$\frac{\ln\left(\frac{D+d(1-1)}{D+(d-1)(1-1)}\right)}{\ln(1)} = \frac{\ln\left(\frac{D+0}{D+0}\right)}{\ln(1)} = \frac{\ln(1)}{\ln(1)} = \frac{0}{0}$$

$$\lim_{F \rightarrow 1} \frac{\ln\left(\frac{D+d(F-1)}{D+(d-1)(F-1)}\right)}{\ln(F)} =$$

$$\lim_{F \rightarrow 1} \frac{\ln(D + d(F - 1)) - \ln(D + (d - 1)(F - 1))}{\ln(F)} =$$

Applying L'Hopital's Rule and differentiating with respect to  $F$ , we obtain:

$$\lim_{F \rightarrow 1} \frac{\frac{d}{(D + d(F - 1))} - \frac{(d - 1)}{(D + (d - 1)(F - 1))}}{1/F} =$$

Direct insertion of 1 for F yields:

$$\begin{aligned} & \frac{\frac{d}{(D+d(1-1))} - \frac{(d-1)}{(D+(d-1)(1-1))}}{1/1} = \\ & = \frac{d}{(D+d(0))} - \frac{(d-1)}{(D+(d-1)(0))} = \frac{d}{D} - \frac{(d-1)}{D} = \frac{1}{D} \end{aligned}$$

This latter proof is about the limit as F approaches 1, while the former proof is about F actually attaining the exact value of 1. Surely one cannot take for granted that the limit of  $f(x)$  as  $x$  approaches  $K$  equals  $f(K)$  since a discontinuity is a distinct possibility. Therefore on the face of it, the two proofs which yield the same  $1/D$  result and complement each other are actually two distinct and necessary results. Yet, as Andrews further points out, both the numerator and the denominator are analytic around  $F = 1$  and L'Hopital's rule guarantees that each had a first-order zero at  $F = 1$ . Consequently their quotient is analytic around  $F = 1$  and its value at  $F = 1$  is precisely the limit that was calculated. Hence the resultant bin equality of  $1/D$  in his proof applies not only to the limit as F approaches 1, but also exactly at that point where F actually equals 1. This renders Andrews' proof not only more straightforward, but also entirely sufficient. Experimentations with two alternative versions of proofs with the aid of L'Hopital's rule by letting F approach 1 first and then letting N approach infinity have ended in decisive failures.

## 9-BIN SYSTEMS WITH $F=10$ ON REAL DATA ALL YIELD $\text{LOG}_{\text{TEN}}(1+1/d)$

---

Significant confirmation of the general bin theory developed here is gotten by way of examining multiple real-life logarithmic data sets and abstract distributions under a nine-bin system having inflation factor  $F = 10$ . Results are then compared to the logarithmic distribution  $\text{LOG}_{10}(1+1/d)$  and a remarkable fit is found! Although such a bin system may appear identical to our number system and digits, as if imitating them, this is clearly not the case for two reasons: (I) such a bin system may start at any point including 0 (as long as it's not too far from the origin), (II) the width of the first bin has a finite value and it may be of a relatively substantial size (within a limit, not too large). Consequently, and most significantly, the bins are not at all aligned and coordinated on the digital marks of our number system. These digital marks are totally insignificant in measuring relative quantities and their absence here has no effect on results whatsoever, which are  $\text{LOG}_{10}(1+1/d)$  just the same. This should be considered as a decisive proof that  $\text{LOG}_{10}(1+1/d)$  is all about relative quantities for the most part, and that its digital application is but a minor event in the much larger quantitative drama.

The table in Fig. 7.20 depicts results from the same real data and distributions of Figs. 7.13 and 7.14 viewed through the lens of a nine-bin system with  $F = 10$ , starting at 0.033 and having an initial width 0.07. The fit into  $\text{LOG}_{10}(1+1/d)$  is quite satisfactory! The very refined and narrow width start of 0.07 and the positioning of the beginning of the whole bin scheme quite near the origin at 0.033 are two essential features contributing to the 'success' of the results in terms of closeness to the logarithmic. The table in Fig. 7.21 depicts results from these real data and distributions viewed through another refined lens of a nine-bin system with  $F = 10$ , starting at 0.06 and having an initial width of 0.027.

A few other empirical tests of these nine data sets using a variety of other small values for the initial width  $w$  and low starting points  $S$  near the origin yield almost identical bin results closely fitting  $\text{LOG}_{10}(1+1/d)$ . These results certainly endow  $\text{LOG}_{10}(1+1/d)$  by far more universal attribute than the one given to it by

Benford via its mere digital interpretation in BL. Here, we truly encounter ‘Benford’s Law’ (30.1%, 17.6%, ... , 4.6%) in its most general form — without digits, free and independent of any number system whatsoever — and this is so for all real-life data and abstract distributions that are known to be ‘logarithmic’. **It must be emphasized that these two bin schemes as well as numerous other ones constructed here have got nothing to do with digits; that no coordination between bin alignment on the x-axis and significant digits exists; and that each bin within each cycle contains a variety of significant first digits mixed in, while resultant overall bin proportion is almost exactly  $\text{LOG}_{10}(1 + 1/d)$ !** Moreover, such  $\text{LOG}_{10}(1 + 1/d)$  empirical results are theoretically supported by the corresponding bin scheme constructed over the generic  $k/x$  distribution, which in and of itself has got nothing to do with digits nor with any possible subterranean existence of some number system imposed below on the x-axis.

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F	Bin G	Bin H	Bin I
Time Between Earthquakes	29.2%	17.8%	13.3%	10.5%	8.1%	6.6%	5.6%	4.8%	4.2%
USA Population Centers	30.3%	17.3%	12.8%	9.7%	7.3%	6.8%	5.7%	5.3%	4.7%
LOG Sym. Triangular (1, 3, 5)	29.6%	17.9%	12.5%	9.6%	8.0%	6.8%	5.8%	5.0%	4.8%
k/x over (1, 1000000)	30.2%	17.6%	12.3%	9.9%	7.8%	6.8%	5.9%	5.0%	4.5%
Exp. Growth, B=1.5, F=1.01	27.4%	18.0%	13.0%	10.1%	8.2%	7.0%	6.0%	5.4%	4.8%
Lognormal, Loc=5, Shape=1	30.7%	18.4%	12.6%	9.3%	7.6%	6.3%	5.6%	5.0%	4.5%
Lognormal, Loc=9.3, Shape=1.7	30.4%	17.4%	12.4%	9.6%	7.9%	6.8%	5.8%	5.2%	4.5%
Varied Data - Hill's Model	32.4%	16.4%	11.2%	8.9%	8.6%	6.1%	6.3%	5.5%	4.6%
Chain U(U(U(U(0, 5666))))))	30.7%	17.3%	12.0%	9.6%	8.1%	6.4%	6.0%	5.6%	4.4%
General R.Q. Law D=9 F=10	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

Figure 7.20 9 Bins F = 10 Yields Nearly  $\text{LOG}_{10}(1+1/d)$  (Start = 0.033 Width = 0.07)

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F	Bin G	Bin H	Bin I
Time Between Earthquakes	31.1%	16.6%	11.9%	9.5%	8.0%	6.6%	6.1%	5.3%	5.0%
USA Population Centers	29.3%	18.3%	12.7%	9.5%	7.6%	6.6%	5.9%	5.3%	4.7%
LOG Sym. Triangular (1, 3, 5)	30.3%	17.5%	12.2%	9.7%	8.1%	6.7%	6.0%	5.2%	4.5%
k/x over (1, 1000000)	30.4%	17.3%	13.0%	9.7%	7.7%	6.8%	5.8%	4.8%	4.5%
Exp. Growth, B=1.5, F=1.01	29.4%	17.2%	12.2%	9.5%	8.1%	7.1%	6.2%	5.4%	4.9%
Lognormal, Loc=5, Shape=1	28.9%	17.4%	12.4%	10.3%	8.3%	6.9%	6.1%	5.2%	4.6%
Lognormal, Loc=9.3, Shape=1.7	29.8%	18.0%	12.5%	9.8%	7.8%	6.3%	5.9%	5.0%	4.7%
Varied Data - Hill's Model	29.5%	19.2%	14.6%	9.1%	7.0%	5.7%	5.9%	4.7%	4.3%
Chain U(U(U(U(0, 5666))))))	30.0%	18.3%	12.6%	9.9%	7.7%	6.7%	5.5%	4.8%	4.5%
General R.Q. Law D=9 F=10	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

Figure 7.21 9 Bins F = 10 Yields Nearly  $\text{LOG}_{10}(1+1/d)$  (Start = 0.06 Width = 0.027)

## BIN SYSTEMS NEED TO START NEAR ORIGIN WITH SMALL INITIAL WIDTH

The table in Fig. 7.22 depicts results from the same list of nine real data sets and distributions as in the previous chapter, viewed through a slightly different bin system prism of  $D = 9$  with  $F = 10$  starting at 1.5 and having an initial width 3. The fit into  $\text{LOG}_{10}(1+1/d)$  is still mostly satisfactory, but not as satisfactory as in the earlier schemes of Figs. 7.20 and 7.21 which come with a much more refined width and start. Worsening results are seen in Fig. 7.23, where the starting point is pushed farther to 5, having an even wider width of 33, since such a bin scheme is considered to be too crude to measure occurrences of relative quantities. When the initial width is way too wide and inclusive, such as seen in Fig. 7.24 which starts at 5, having an initial width value of 311, results are off, and often are heavily skewed in favor of low bins.

The explanation for all this is straightforward, namely that whenever the initial bin width is quite large relative to the spread of the data, the first bin on the first cycle captures most or a big part of overall data, leaving much less for the second bin adjacent to it. The two distinct bin system outlines shown in Figs. 7.25 and 7.26 represent two distinct attempts at measuring the fall of the same

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F	Bin G	Bin H	Bin I
Time Between Earthquakes	30.4%	16.8%	11.8%	9.7%	8.0%	6.8%	5.9%	5.6%	4.9%
USA Population Centers	29.5%	18.3%	12.4%	9.6%	7.6%	6.7%	6.0%	5.2%	4.6%
LOG Sym. Triangular (1, 3, 5)	30.3%	17.5%	12.2%	9.8%	7.8%	6.9%	5.9%	5.1%	4.6%
k/x over (1, 1000000)	34.0%	18.9%	12.1%	8.5%	6.8%	6.1%	4.8%	4.4%	4.3%
Exp. Growth, B=1.5, F=1.01	31.0%	17.6%	12.4%	9.5%	7.7%	6.6%	5.7%	5.0%	4.4%
Lognormal, Loc=5, Shape=1	28.9%	17.3%	12.9%	10.1%	8.3%	7.0%	5.8%	5.2%	4.5%
Lognormal, Loc=9.3, Shape=1.7	30.1%	17.9%	12.5%	9.7%	7.6%	6.8%	5.7%	5.3%	4.5%
Varied Data - Hill's Model	32.4%	19.7%	12.7%	9.0%	7.1%	6.2%	5.0%	4.4%	3.5%
Chain U(U(U(U(U(0, 5666))))))	32.7%	18.8%	12.4%	9.3%	7.5%	6.2%	4.9%	4.3%	4.0%
General R.Q. Law D=9 F=10	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

Figure 7.22 9 Bins  $F = 10$  Deviates Somewhat from  $\text{LOG}_{10}(1+1/d)$  (Start = 1.5 Width = 3)

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F	Bin G	Bin H	Bin I
Time Between Earthquakes	30.1%	16.7%	12.3%	10.1%	7.6%	7.0%	5.9%	5.5%	4.9%
USA Population Centers	29.7%	17.8%	12.5%	9.4%	7.9%	6.7%	6.1%	5.2%	4.6%
LOG Sym. Triangular (1, 3, 5)	30.1%	18.8%	12.6%	9.9%	7.9%	6.6%	5.4%	4.7%	4.0%
k/x over (1, 1000000)	39.3%	18.8%	10.5%	7.7%	6.6%	4.7%	4.5%	4.4%	3.5%
Exp. Growth, B=1.5, F=1.01	33.3%	17.3%	12.0%	9.2%	7.4%	6.3%	5.4%	4.8%	4.3%
Lognormal, Loc=5, Shape=1	25.1%	19.0%	14.4%	11.5%	8.9%	7.1%	5.6%	4.6%	3.7%
Lognormal, Loc=9.3, Shape=1.7	30.1%	17.9%	12.5%	9.6%	7.7%	6.8%	5.8%	5.0%	4.6%
Varied Data - Hill's Model	38.6%	19.3%	12.3%	7.5%	5.8%	5.5%	4.4%	3.5%	3.0%
Chain U(U(U(U(U(0, 5666))))))	45.0%	18.4%	10.6%	7.5%	5.7%	4.4%	3.3%	2.7%	2.5%
General R.Q. Law D=9 F=10	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

Figure 7.23 9 Bins F = 10 Deviates Markedly from  $LOG_{10}(1+1/d)$  (Start = 5 Width = 33)

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F	Bin G	Bin H	Bin I
Time Between Earthquakes	35.9%	17.7%	12.3%	9.0%	7.5%	5.6%	4.7%	3.8%	3.4%
USA Population Centers	36.1%	21.3%	12.9%	8.8%	6.1%	4.8%	4.1%	3.3%	2.6%
LOG Sym. Triangular (1, 3, 5)	42.0%	18.4%	11.3%	8.0%	6.1%	4.7%	3.8%	3.1%	2.6%
k/x over (1, 1000000)	51.3%	15.8%	8.8%	6.1%	4.8%	4.3%	2.8%	3.2%	2.9%
Exp. Growth, B=1.5, F=1.01	39.3%	15.8%	11.0%	8.4%	6.8%	5.7%	4.9%	4.4%	3.8%
Lognormal, Loc=5, Shape=1	77.6%	14.9%	4.3%	1.7%	0.8%	0.4%	0.2%	0.1%	0.1%
Lognormal, Loc=9.3, Shape=1.7	29.9%	18.2%	12.7%	9.8%	7.6%	6.7%	5.8%	4.9%	4.4%
Varied Data - Hill's Model	53.2%	13.8%	9.4%	6.1%	4.7%	3.6%	4.0%	3.0%	2.3%
Chain U(U(U(U(U(0, 5666))))))	80.3%	11.4%	4.3%	1.8%	1.0%	0.6%	0.3%	0.2%	0.1%
General R.Q. Law D=9 F=10	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

Figure 7.24 9 Bins F = 10 Greatly Deviates from  $LOG_{10}(1+1/d)$  (Start = 5 Width = 311)

data density. This clearly demonstrates the rationale behind the need to cast a refined and thin net of bin system in order to properly measure fall in density and occurrences of relative quantities. The initial thick bin system in Fig. 7.25 does not even complete a single cycle before termination of the data is reached, and thus an exaggerated portion of data falls within the first bin A. Moreover, it doesn't do justice to bin C as it deprives it almost of any data. In Fig. 7.26 on the other hand, the more refined bin system manages to turn at least three full cycles before termination of the data is reached. Another obvious requirement in general is for the bins to start at or very near the origin so as to capture all the data. Even when there is no data at all near the origin, all bin schemes should be standardized and united in starting at the same point (which might as well be zero), and should have the same initial (small) width  $w$ , otherwise there can be no consistent and steady law to observe and state, one that can be applied across all logarithmic data sets.

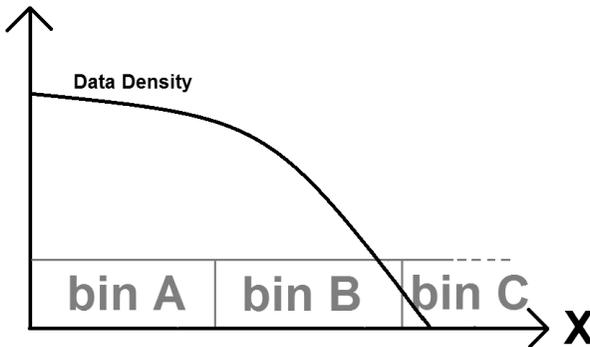


Figure 7.25 Too Thick an Initial Bin Width Distorts Results (Over-skewness)

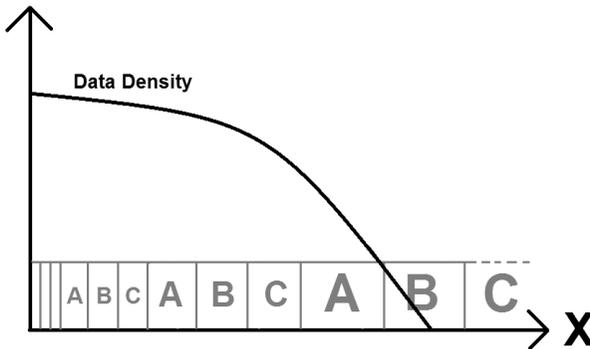


Figure 7.26 Refined and Thin Initial Bin Width is Needed for Correct Results

One crucial aspect of all bin schemes and bin laws is that empirically exact values of the initial width  $w$  and especially the starting point  $S$  do not matter for the most part in logarithmic data sets, as long as  $S$  and  $w$  are made small (relative to the spread of the data sets in question). In other words, empirical bin results for logarithmic data sets are consistent and are  $w$ -invariant as well as  $S$ -invariant. Separately, for  $w$  and for  $S$ , this can also be said about the infinitely expanded bins for the generic  $k/x$  case, although the model is restricted to  $w = S$ . For the group of nine data sets above, any combination of  $w$  and  $S$  less than 0.1 gave almost the same result (given fixed values of  $F$  and  $D$ ). Only when values of  $S$  and  $w$  were made progressively larger did bin proportions become increasingly more and more

dependent on them. For low values of  $S$  and  $w$ , the only factors in bin laws are: (I) number of bins  $D$ , and (II) inflation factor  $F$ . It is advisable in any case to universally restrict  $S < 0.001$  and  $w < 0.001$  in all bin schemes. An arbitrarily self-imposed standardization rule such as  $S = 0$  and  $w = 0.0005$ , for example, in all bin schemes could facilitate consistent comparisons in the field. Yet, when the data set in question falls mostly on, say, the interval  $(0, 0.0005)$ , the above standardization rule is not sufficient at all.

It must be noted that there is nothing in the particular structure of the bin system in Fig. 7.26 that renders it universally refined. Refinement is relative to the data curve in question. The same bin system of Fig. 7.26 is shown again in Fig. 7.27 where it is considered too thick and crude for Data A, but refined and thin enough for Data B. Our bin system needs to imitate, borrow, and learn from Benford's Law how to infinitely refine its width towards the origin. This would serve two purposes: (I) standardizing the bin structure irrespective of data spread, and (II) making sure that bins are refined enough (at least near the origin). Alas, this is still not a cure-for-all solution as seen in the annoying example of Fig. 7.28 where data starts far from the origin around where bins have gotten already thick and crude and terminates soon afterwards, spanning only one bin cycle. Typically in cases such as the data curve in Fig. 7.28, the eager, impulsive, and inexperienced data analysts may have the temptation to start the same bin structure just to the left of the minimum value so that many multiple cycles will be covered. Such an adjustment though would ruin standardization of a common start at, say, the origin, and should be rejected. The data analysts can't and should not perform different bin systems for different data sets. If each bin scheme is tailor-made for each particular

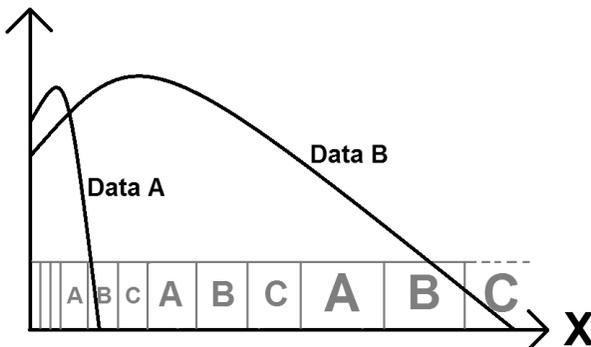
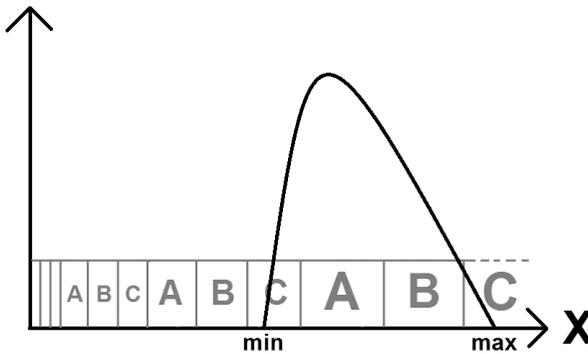


Figure 7.27 Refinement Being Relative to Data Spread



**Figure 7.28** Refinement is Already Lost When Data Starts Late and Terminates Fast

data set, no universal law could be stated. The data set in Fig. 7.28 is strongly suspected of being ‘non-logarithmic’ by nature, and as such can never hope to join other logarithmic data sets with regards to bin proportions. The difficulty of the data in Fig. 7.28 is akin to the situation in Benford’s Law where data suffers from small order of magnitude.

**In order to formally define the bin model for data sets with only non-zero positive numbers, the starting point  $S$  is made to be at the origin and the initial bin width  $w$  is made infinitesimally small by defining it as a limiting process where  $w$  approaches zero.**

In order to demonstrate how changes in the values of  $w$  or  $S$  affect bin results, two distinct bin schemes are performed: one for the logarithmic data set pertaining to time between 2012 earthquakes, and another for the non-logarithmic U.S. county area data set. Figure 7.29 depicts a variety of bin results for a 5-bin scheme with  $F = 3$  on the earthquake data. Clearly all of these changes in the values of  $w$  and  $S$  do not significantly affect results, as long as both  $w$  and  $S$  remain relatively small. Remarkably, results seem to be almost totally independent of  $S$ ! Figure 7.30 depicts various bin results for a 3-bin scheme with  $F = 7$  on the U.S. county area data. Since the data set is not logarithmic, tiny changes in  $w$  here greatly affect bin results! In other words, non-logarithmic data is highly sensitive to small changes in initial bin width  $w$ . Remarkably, even for this non-logarithmic data set, results seem to be almost totally independent of  $S$  (so long as  $S$  remains sufficiently small).

Starting	Initial Width	Bin A	Bin B	Bin C	Bin C	Bin D
0	0.000000005	30.4%	22.5%	18.3%	15.3%	13.5%
0	0.002	31.0%	22.9%	17.9%	14.9%	13.2%
0	0.007	31.2%	22.3%	18.0%	15.6%	13.0%
0	0.033	30.0%	23.3%	18.6%	15.2%	12.9%
0	0.822	30.2%	23.0%	18.6%	15.1%	13.0%
0	3.8	31.1%	23.2%	18.0%	14.9%	12.8%
0.006	0.3	30.2%	23.1%	18.7%	15.1%	12.9%
0.043	0.3	30.2%	23.1%	18.7%	15.1%	12.9%
0.077	0.3	30.2%	23.1%	18.7%	15.1%	12.9%
0.257	0.3	30.2%	23.2%	18.7%	15.1%	12.8%
2	0.3	30.2%	23.2%	18.6%	15.1%	12.9%
3	0.3	30.2%	23.2%	18.7%	15.2%	12.8%
<b>Law of Relative Quantities</b>		<b>30.6%</b>	<b>22.9%</b>	<b>18.3%</b>	<b>15.2%</b>	<b>13.0%</b>

Figure 7.29 S and w Effects on Time between 2012 Earthquake Data Set,  $D = 5 F = 3$

Starting	Initial Width	Bin A	Bin B	Bin D
0.0043	0.000002	68.4%	21.9%	9.7%
0.0043	0.00006	46.4%	37.1%	16.5%
0.0043	0.03	66.1%	25.0%	9.0%
0.0043	0.5	51.2%	22.6%	26.2%
0.0043	2	73.9%	14.9%	11.2%
0.0043	5	40.3%	37.9%	21.7%
0.006	0.389	64.8%	15.2%	20.0%
0.029	0.389	64.8%	15.2%	20.0%
0.706	0.389	64.7%	15.3%	20.0%
0.0034	0.0143	40.5%	37.1%	22.4%
0.005	0.0143	40.5%	37.1%	22.4%
0.76	0.0143	40.5%	37.2%	22.3%
2	0.0143	40.8%	37.0%	22.2%
3	0.0143	40.9%	37.0%	22.1%
<b>Law Relative Quantities</b>		<b>56.5%</b>	<b>26.3%</b>	<b>17.3%</b>

Figure 7.30 S and w Effects on U.S. County Area Data Set,  $D = 3 F = 7$

## ACTUAL OR DEGREE OF COMPLIANCE MAY BE BIN- AND BASE-VARIANT

---

The concern raised in the previous chapter, such as the difficulty in the data of Fig. 7.28, strongly reminds Earthlings of their own very similar preoccupation within the context of digital Benford's Law, namely Order Of Magnitude (OOM), which may be considered as one of the most important factors in observing the law, but which must also involve at least some fall in the density with a tail to the right for logarithmic behavior to be observed. In Chapter 46 it was shown that **OOM is scale-invariant**. Yet OOM is not base-invariant. In order to demonstrate this crucial fact we shall refer again to Fig. 7.28, now conveniently assumed to represent physical data on planet Clarikia and clearly observable by two distinct planets inhabited by civilizations using positional number system with the distinct B1 and B2 bases. Calculating OOM for base B1, we get:

$$\text{OOM}_{B_1} = \text{LOG}_{B_1}(\text{max}) - \text{LOG}_{B_1}(\text{min}) = \text{LOG}_{B_1}(\text{max}/\text{min}).$$

Calculating OOM for base B2, we get:

$$\text{OOM}_{B_2} = \text{LOG}_{B_2}(\text{max}) - \text{LOG}_{B_2}(\text{min}) = \text{LOG}_{B_2}(\text{max}/\text{min}).$$

Applying the logarithmic identity  $\text{LOG}_A X = \text{LOG}_B X / \text{LOG}_B A$  we obtain:

$$\begin{aligned} \text{OOM}_{B_2} &= \text{LOG}_{B_2}(\text{max}/\text{min}) = \text{LOG}_{B_1}(\text{max}/\text{min}) / \text{LOG}_{B_1}(B_2) = \\ &= \text{OOM}_{B_1} / \text{LOG}_{B_1}(B_2) \end{aligned}$$

And finally we obtain the expression  $\text{OOM}_{B_2} = \text{OOM}_{B_1} / \text{LOG}_{B_1}(B_2)$ .

For example, if OOM in base 10 for a given data set is 3, then OOM in base 4 with respect to the same data set is  $\text{OOM}_4 = \text{OOM}_{10} / \text{LOG}_{10}(4) = 3 / 0.602 = 5$ . This state of affairs is rather peculiar. While both planets passionately believe in and apply digital Benford's Law, yet they view that same Clarikian physical data set differently with regards to OOM and (on the face of it) perhaps even with regards to compliance. Data compliance or degree of deviation then appears to be relative to the observing base, not absolute, even though the two planets keep staring at the same physical phenomenon! This result demonstrates that **OOM is not**

**base-invariant**, but is it possible that it also demonstrates that different planets with distinct bases adopt totally different views and conclusions about conformity and degree of deviation from Benford's Law regarding any physical data set, in spite of the well-known and accepted fact that **Benford's Law is base-invariant**? This is not really a paradox in any way, since the general expression of proportions  $\text{LOG}_{\text{BASE}}(1 + 1/d)$  is applied to perfectly logarithmic data, hence the law is called 'base-invariant'. Yet, non-logarithmic data set may deviate from the law to a different degree depending on the base-point-of-view. Logarithmic data sets do have indeed different OOM values for different planets with distinct bases, but since those two OOM values are sufficiently large on both planets overshooting their respective logarithmic OOM thresholds, both agree that the data is indeed logarithmic, regardless; it nicely fits  $\text{LOG}_H(1 + 1/d)$  on the planet with base H, and it nicely fits  $\text{LOG}_K(1 + 1/d)$  on the planet with base K.

The OOM cutoff point for compliance is different for different bases. For example, base 4 demands higher OOM value for compliance with the law than what base 10 would demand. If  $\text{OOM}_{10} = 3$  is demanded on Earth for base 10 as the threshold, then  $\text{OOM}_B = 3/\text{LOG}_{10}(B)$  is demanded on any other planet utilizing base B. For example, base 4 demands  $\text{OOM}_4 = 5$  as the minimum value for logarithmic behavior; while base 31 demands only  $\text{OOM}_{31} = 2$  as the minimum value. Physically what we demand from the data is that  $(\text{max}/\text{min}) > (\approx 1000)$ , and this requirement is quantitative, namely number system invariant; while we allow each planet to calculate its own  $\text{LOG}_B(\text{required max}/\text{min})$  namely  $\text{LOG}_B(\approx 1000)$  to its heart content, to express the requirement in its own way and familiar numerical style.

As an example of how OOM and degree of non-conformity with digital Benford's Law is base-dependent, U.S. county area data set shall be re-considered. The data set is decisively non-Benford because of its insufficient OOM. In order to avoid the misguided influence not only from outliers but also from the edges in the calculation of the measure, only, say, the middle 80% of overall data should be considered, and Order of Magnitude of Variability  $\text{OMV}_{10}$  is calculated. Here  $Q_{10\%}$  is 286,  $Q_{90\%}$  is 1843, and consequently the middle 80% bulk of the data lies in the interval (286, 1843). Therefore  $\text{OMV}_{10} = \text{LOG}_{10}(Q_{90\%}/Q_{10\%}) = \text{LOG}_{10}(1843/286) = 0.81$ . Since  $\text{OMV}_{10} = 0.81$  (considered less than the usual 2.5 units required in base 10 for logarithmic behavior) the data does not conform. First digits for the U.S. county area data as well as Benford's Law in base 10 are shown for comparison:

Data set = {16.2%, 10.0%, 10.7%, 15.8%, 15.2%, 10.4%, 8.6%, 7.1%, 5.9%}  
 Benford = {30.1%, 17.6%, 12.5%, 9.7%, 7.9%, 6.7%, 5.8%, 5.1%, 4.6%}

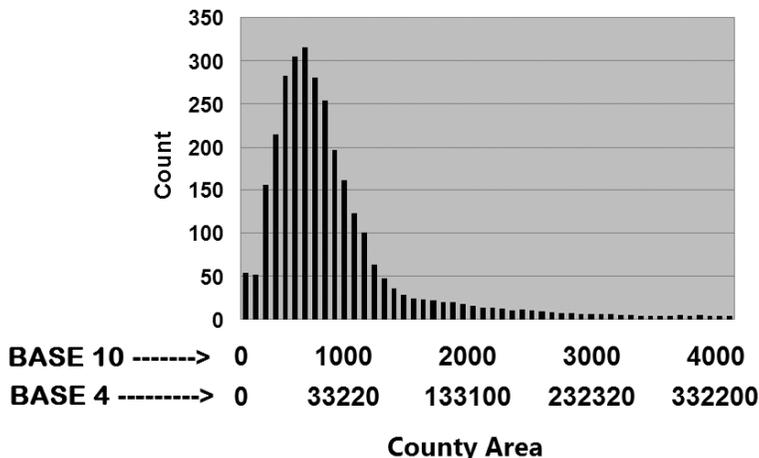


Figure 7.31 Histogram of U.S. County Area Viewed via Base 10 and Base 4

When all area values are transformed into base 4, deviations of observed first digits from Benford's Law base 4 of  $\text{LOG}_4(1+1/d)$  may be seen to be milder perhaps:

Data set = {43.2%, 35.6%, 21.2%}

Benford = {50.0%, 29.2%, 20.8%}

Calculating  $\text{OMV}_4$  we obtain the value  $\text{LOG}_4(1843/286) = 1.34$ , therefore  $\text{OMV}_4 = 1.34$  which is larger than  $\text{OMV}_{10}$ . Could this explain why in base 4 deviations may appear a bit milder? Not if one recalls that in base 4 we demand higher OOM in general! Figure 7.31 depicts the histogram of U.S. county area viewed by way of base 10 and base 4 markings on the x-axis. Outliers extend much further to the right until approximately 25,000, but are ignored as they are few and far between.

Figure 7.32 depicts the bulk of U.S. county area histogram, narrowing the focus a bit, and clumsily (and unsuccessfully perhaps) attempts to demonstrate why milder deviations from base-4-law might be expected as compared with deviations from base-10-law. Evidently, integral powers of 10 and integral powers of 4 form very different marking patterns on the x-axis as they are of different cyclical lengths. Base 4 gives rise to more concentrated (frequent) such markings. Hence, for any given real-life physical data whose histogram is fixed on the x-axis independently of any number system and bases, OOM is higher in lower bases but this fact does not really explain why it might be considered closer to the logarithmic.

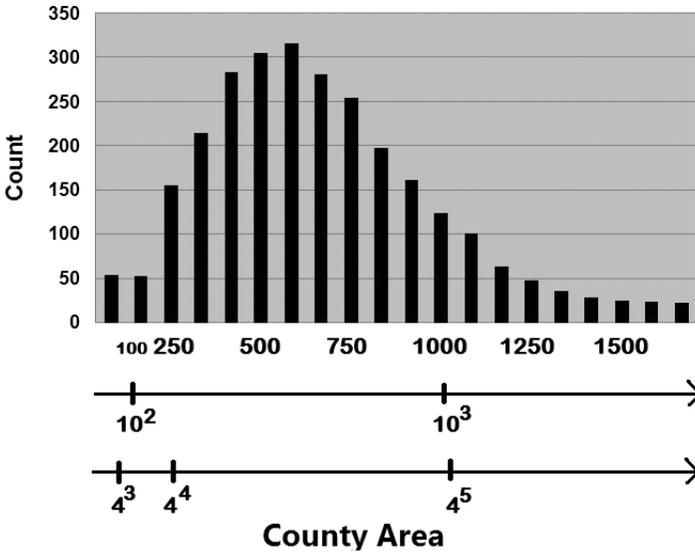


Figure 7.32 Shorter Digital Cycle in Base 4 — Longer in Base 10 — U.S. County Area

Benfordian analysts and mathematicians firmly believe that there exists only one digital law, regardless of base, and to support their misguided claim they point to the fact that one single analytical algebraic expression captures the whole phenomenon, as in  $\text{LOG}_{\text{BASE}}(1+1/d)$ . This view is supported and propagated by the Federation which has proclaimed inter-base Benfordian unity as a divine sign of galactic commonality in order to promote inter-planetary peace, but experienced data analysts and galactic fraud detectors traveling on all four arms and far beyond and having to deal with so many different planetary customs, number systems and bases, believe that there are numerous distinct laws for all practical matter, a particular law for each base in use. When mathematicians and Federation officials are asked to compare two deviating data sets from two different planets using distinct bases in order to decide on degree of severity they are at a complete loss. Applying SSD measure to compare deviations of two data sets should ideally be done only when both sets belong to the same base and hence aiming at the same numerical set (vector) of proportions, i.e. the same law. Even more embarrassing for them is the fact that anomalous/rebellious exponential growth series relative to base X with integers L and K and rational number  $\text{LOG}_X(F) = L/K$  is suddenly totally obedient and normal in another base Y and now  $\text{LOG}_Y(F) \neq L/K$  and quite irrational. Even the requirement that normal but short exponential (low) growth

Benford 10	County Area	(Ben - Obs) <sup>2</sup>
30.1%	16.2%	192.6
17.6%	10.0%	57.6
12.5%	10.7%	3.4
9.7%	15.8%	37.9
7.9%	15.2%	53.1
6.7%	10.4%	13.8
5.8%	8.6%	7.6
5.1%	7.1%	4.0
4.6%	5.9%	1.9

SSD = 371.8

Benford 4	County Area	(Ben - Obs) <sup>2</sup>
50.0%	43.2%	46.2
29.2%	35.6%	41.0
20.8%	21.2%	0.2

SSD = 87.4

**Figure 7.33** SSD is Lower in Base 4 than in Base 10

series needs to be bordered by integral powers of the base is base-dependent. The same confusing fate befalls  $k/x$  over an integral exponent difference which involves the base and is judged differently on the issue of digital Benfordness on planets with distinct bases. For example,  $k/x$  defined over (10, 100) is logarithmic in base 10, and decisively non-logarithmic in base 4 (general proof of this is given below). These few cases of ('short')  $k/x$  and exponential growth series show in extreme generality that their logarithmic property is not absolute but rather relative to the base. Our U.S. county area data set faces a similar dilemma, namely within which base system should it be considered more deviant or less deviant in the context of digital Benford, base 4 or base 10? The question itself is perhaps mathematically meaningless since 'deviation' hasn't been inter-base defined. The table in Fig. 7.33 depicts SSD calculations for U.S. county area in both bases. Most laymen upon staring at Fig. 7.33 instinctively point to the base 4 perspective as being the one where data is closer to its theoretical proportions. To back up this intuition, the suggested inter-base measure of deviation is simply the average (among all the relevant digits) of all percent deviation squares within each base, namely the ratio  $\Sigma(O_i - B_i)^2 / [\text{number of } i \text{ digits}]$ , or  $SSD / [\text{number of digits}]$ , which for the U.S. county area data set points to  $(371.8) / (9)$  or **41.3** as the

average in base 10, and  $(87.4)/(3)$  or **29.1** as the lower average in base 4. It would not do justice perhaps to simply compare SSD values since there are more digits in base 10 than in base 4. In conclusion, for this U.S. county area data set: since base 10 yields higher SSD as well as higher average deviation squares per digit in comparison with base 4, deviation from Benford is said to be greater in base 10 than in base 4.

Let us prove the above assertion regarding dependency of  $k/x$  defined over  $(A, B)$  on base with regards to logarithmic behavior. Given that the distribution is logarithmic in base  $X$ , namely that the exponent difference is an integral number, therefore:

$$\text{LOG}_X(B) - \text{LOG}_X(A) = \text{INTEGER}$$

Let us consider the same  $k/x$  distribution on  $(A, B)$  viewed from base  $Y$  perspective. From the above equation we get:

$$\text{LOG}_Y(B)/\text{LOG}_Y(X) - \text{LOG}_Y(A)/\text{LOG}_Y(X) = \text{INTEGER}$$

$$[\text{LOG}_Y(B) - \text{LOG}_Y(A)]/\text{LOG}_Y(X) = \text{INTEGER}$$

$$[\text{LOG}_Y(B) - \text{LOG}_Y(A)] = \text{INTEGER} * \text{LOG}_Y(X)$$

$$[\text{exponent difference in base } Y] = \text{INTEGER} * \text{LOG}_Y(X)$$

Hence, exponent difference in the new base  $Y$  could also be an integer if and only if  $\text{LOG}_Y(X)$  is an integer or a rational fraction which yields some integer when multiplied by  $\text{INTEGER}$ , which is rare. For example,  $\text{LOG}_{10}(100)$  is an integer.  $\text{LOG}_3(9)$  is also an integer. But  $\text{LOG}_{10}(16)$  or  $\text{LOG}_4(10)$  are not integers. Worse yet, they are irrational numbers, and any irrational number multiplied by any integer can never hope to metamorphose into an integer by the mere act of such multiplication.

How do we align  $k/x$  defined within a narrow  $(A, B)$  interval over the  $x$ -axis in the context of bin schemes so as to obtain that equivalent 'integral exponent difference' and the resultant 'logarithmic' behavior? It must be acknowledged that neither  $(D + 1)$  is the base in the usual sense, nor inflation factor  $F$  of course. It is necessary to give the term more general significance with wider applicability. The meaning of 'integral exponent difference' in its purest form, encompassing bin systems as well, signifies having points  $A$  and  $B$  aligned on the  $x$ -axis in such a way that exactly a whole bin cycle has been turned, or two whole cycles, or  $N$  whole cycles. In other words, that the left corner of the first bin in the first cycle is exactly at  $A$ , and that the right corner of the last bin in the last cycle is exactly at  $B$ . As a brief empirical check on this conceptual assertion, Scheme A bin system

of Fig. 7.13, namely  $D = 4$ ,  $F = 8$ ,  $S = 0$ ,  $W = 0.0008$ , is cast over the x-axis and the distribution  $k/x$  defined over  $(0.0032, 0.0288)$  is compared to the General Law. The interval  $(0.0032, 0.0288)$  is exactly the second whole cycle of the grid of the bin scheme itself, namely the interval  $(4*0.0008, 4*0.0008+4*8*0.0008)$ . Empirical results show strong conformity with that particular quantitative ‘bin law’ of Scheme A. Resultant bin proportions for this  $k/x$  distribution yield  $\{49.2\%, 23.7\%, 15.8\%, 11.3\%\}$  and are nicely compared to the General Law of  $\{48.6\%, 23.7\%, 15.8\%, 11.9\%\}$ . Surely the whole narrow interval is allowed to shift to the left or to the right without adversely affecting results, just as was seen in digital Benford’s Law for, say,  $k/x$  over  $(20, 200)$ . Such a shift would let the rightmost point terminate in the same cyclical phase as that of the leftmost point. Needless to say, the existence of a universal or absolute concept such as ‘finite logarithmic  $k/x$ ’ in the context of bin schemes is not so obvious, since compliance seems to be relative to the bin structure in question and to the types of particular cycles turning, just as the existence of an absolute concept of finite logarithmic  $k/x$  within the context of digital Benford’s Law irrespective of the base in the number system poses a challenge. Nonetheless,  $k/x$  distribution defined over an extremely large x-axis range may be construed as being universally logarithmic in the approximate, with respect to all possible bin schemes having any  $D$  and  $F$  values, as well as with respect to all bases in digital Benford’s Law, as shall be discussed in the next chapter.

## CORRESPONDENCE IN DATA CLASSIFICATION BETWEEN BIN SYSTEMS AND BL

---

U.S. county area data set was found to be non-logarithmic in the context of digital Benford's Law considering either base 4 or base 10. But could this data set hope to fit in nicely within other 'logarithmic' data sets in the context of bin schemes? In other words, could a rejected non-logarithmic data set in a digital BL sense find absolution in some bin system, becoming accepted at long last by its 'logarithmic' data peers as equal? The answer is decidedly in the negative. Casting the two bin schemes A and B as in Figs. 7.13 and 7.14 over the x-axis for U.S. county area data set yields the following results:

Scheme A:

$$D = 4 \quad F = 8 \quad S = 0 \quad W = 0.0008$$

Avg. of all 9 real data sets = {48.3%, 23.8%, 15.9%, 12.0%}

U.S. County Area data = {29.2%, 29.2%, 26.3%, 15.3%}

Scheme B:

$$D = 7 \quad F = 3 \quad S = 0 \quad W = 0.0008$$

Avg. of all 9 real data sets = {22.6%, 18.4%, 15.2%, 13.1%, 11.3%, 10.1%, 9.3%}

U.S. County Area data = {26.6%, 18.3%, 15.4%, 10.9%, 9.7%, 10.9%, 8.1%}

Clearly U.S. county area data set has a different bin behavior than the rest of the data sets in the group. This is an extremely crucial and significant fact that must be acknowledged, namely **that [for a data set] being digit-wise logarithmic in the context of Benford's Law and obeying in general all bin systems laws go together. Having non-logarithmic digit configuration corresponds to having non-bin-law bin configuration.** Yet this fact should not be surprising in the least, because the digital scheme developed by Frank Benford is nothing but one particular example of the generic idea of the bin scheme, a singular manifestation of a much larger universe of bin system possibilities. A special one to be sure, due to  $F = D + 1$ , but a bin scheme nonetheless!

In conclusion, being ‘logarithmic’ or ‘non-logarithmic’ is an absolute and universal property of any data set, irrespective of the [number system] base or bin scheme in use. Curiosity compels one to ask then: “What is the basic or the most essential characteristic of a ‘logarithmic’ data set?” Do we need to resort to checking compliance with digital BL or bin schemes laws in order to characterize a given data set as such? Couldn’t we find some more direct way to measure this aspect of the data from its own intrinsic quantitative arrangement and structure, without ‘looking outside’ for compliance with digital or bin laws, and without it desperately seeking approval and acceptance from its data peers? An appealing answer to this dilemma is the statement that logarithmic-ness is the intrinsic property of having an overall (average) decrease in relative quantities in the same rapidity and manner as that of the  $k/x$  generic distribution infinitely expanded. Such an interpretation does not involve fitting a given data set into multiple ‘logarithmic’ data sets group behavior, but rather performing a singular measure of compatibility with  $k/x$ . Perhaps one may sum up the uniqueness and simplicity in the fall of  $k/x$  by pointing to the fact that by doubling  $x$  value we cut the density exactly by a half, hence Allaart’s sum invariance characterization principle since this represents an exact ‘trade-off’. For example, for the distribution  $0.4342945/x$  defined over  $(10, 100)$ , density height or histogram count on  $x = 40$  is exactly half that on  $x = 20$ . This implies that the height of the curve times the  $x$  value below [i.e. ‘sum’] for any infinitesimal sub-interval is a constant everywhere. Such an exact relationship between  $x$  and its quantitative frequency is of course unique to  $k/x$  distribution.

It must be emphasized that such a definition of ‘logarithmic-ness’ utilizes  $k/x$  defined over a ‘truly long range’, namely  $k/x$  defined over  $(\approx a, \approx \infty)$ . For all practical purposes one can generate  $k/x$  data on, say,  $(0.0000287, 6.53 \cdot 10^{13})$  and call it logarithmic. Earthlings would instinctively prefer say  $k/x$  over  $(0.00001, 10^{15})$ , which is just as good, but such a tendency is due to their obsession with digits, number systems, and base 10. It may be a bit shocking, but  $k/x$  defined over  $(1, 10)$  is not logarithmic in this sense! It only complies with digital Benford’s Law when base 10 is used, not when other bases are used, and it certainly does not comply with any bin scheme laws unless they mimic

exactly BL base 10. Short-range  $k/x$  are ‘too sensitive’ to base or bin changes. Long-range  $k/x$  are crude and stable, being base-invariant and bin-scheme-invariant, not in the sense that proportions [laws] are the same in different bases or bin systems; they certainly are not, but in the sense that long-range  $k/x$  obey **all** digital and bin laws regardless. The main challenge for  $k/x$  defined over a short range in obeying digital and bin laws is a perfect match with the measuring cycle, namely that  $k/x$  does not abruptly start or terminate in the middle of a bin/digit cycle which strongly discriminates against some bins/digits, but rather that it starts at the beginning of a cycle and that it terminates at the end of a cycle, or that shifted ranges starts and terminates at the same cyclical phase, as in  $k/x$  over (20, 200). For example, the distribution  $k/x$  defined over (10, 200) abruptly terminates in the middle of the digital cycles  $10^{\text{INTEGER}}$  and thus discriminates against digits 2 to 9, while unfairly favoring digit 1. The distribution  $k/x$  defined over (1, 1000) on the other hand, starts and terminates exactly at the endpoints of the digital cycles, and therefore it obeys BL base 10. For  $k/x$  defined over a very long range, this issue is a minor one, since any such abrupt launch or termination incorporating a partial cycle has a very small effect in the grand scheme of things.

All this is also nicely consistent with the consideration of the General Law representing the limit of the sequence of algebraic expressions of bin systems for  $k/x$  where  $F \neq D + 1$ , since convergence to the limit necessitates the consideration of a ‘very long range’ on the  $x$ -axis in an infinitely expanding cycles of bins! Approximate convergence there is not found in, say, merely five or 10 cycles, but rather it is often obtained after, say, 30 or 100 full cycles, while for some  $D$  and  $F$  combinations it may take several hundred or more cycles just to get somewhat close enough to the true limit. On the other hand, bin schemes casting a refined net from the very origin to measure real data, has a different focus in the opposite direction. Any real positive data surely terminates at some ‘very finite’  $P$  point, while the grid of the bin system measuring (fall in) relative quantities constitutes an **infinite** net of bins emerging from the **left** towards  $P$ . The General Law on the other hand is derived from **infinitely** expanding bins **rightwards** from  $w$  to  $\infty$  for the  $k/x$  distribution. Yet, in spite of this ‘**directional dichotomy**’ both bin proportions of real finite data and the General Law perfectly correspond.

As noted in Chapter 124, Comment III, all the mathematical derivations for the General Law on the relative quantities of  $k/x$  distribution assume equality between the width of the bins in the first cycle and the separation from the

0 origin at launching, as can be seen from the designations 0, w, 2w, 3w, and so forth, on the x-axis in Fig. 7.15. Violation of this equality leads to different results in the finite expansion case, where separation that is longer than the size of the first bin yields greater bin equality, and separation that is shorter than the size of the first bin yields sharper fall and extreme bin inequality. However, when the number of expansions goes to infinity this crucial constraint becomes irrelevant and results are independent of the exact launching point. This fact provides us with one more crucial brick in the whole harmonious and consistent bin edifice, guaranteeing that any bin result of any particular  $k/x$  defined over a very large interval (quantitative-order-of-magnitude-wise), would closely correspond to the theoretical limit of the bin proportions of infinite cycles construction built onto the generic  $k/x$  distribution [the General Law]. For example, bin proportions for  $k/x$  defined over the interval (8, 753,947,289) should almost exactly correspond to the General Law for all value combinations of D and F, and one need not worry about adjusting any separation from origin or bin width size to 8. On the other hand, bin proportions for  $k/x$  defined over, say, (8, 22) should pose a serious challenge in terms of matching the separation and bin width  $w$  to 8, as well as in terms of guaranteeing that the last finite bin cycle terminates somehow exactly at 22. In addition, the unambiguous ‘finiteness’ of any bin structure constructed for  $k/x$  over (8, 22) would impede convergence to the limit of that infinite sequence, thus yielding different results. One should not lose sight though of the fact that it’s ‘quantitative-order of magnitude’ [of the whole range from the leftmost point to the rightmost point] that affects correspondence here, not strictly ‘big’ or ‘small’ values. For example, bin proportions for  $k/x$  over (0.000008, 22) corresponds nicely to the General Law, while results for  $k/x$  over (757,890, 1,323,776) do not even come close.

A measure called Quantitative Order of Magnitude (QTM) shall now be defined in the context of the general law of relative quantities. Since the two previously defined related measures of  $OOM_{BASE} = LOG_{BASE}(max/min)$  &  $OMV_{BASE} = LOG_{BASE}(Q_{90\%}/Q_{10\%})$  assume a [positional] number system and a particular base, it is necessary to construct a purely quantitative measure that would not only be independent of the base in use, but also one that would be totally independent of any number system in and of itself.

**Quantitative Order of Magnitude (QTM) =  $Q_{90\%}/Q_{10\%}$**

Surely one can now express this ratio in a positional number system, Tau Cetian symbols [with difficulties], or Roman Numerals. But all this is irrelevant as the

above expression signifies a primitive and basic concept of quantity that does not necessitate any number system whatsoever. Could logarithmic-ness then be simply reduced to such basic quantitative measure? Certainly not! Logarithmic-ness is by far more complex than the mere spread of data. The table in Fig. 7.34 depicts a large variety of data sets, their QTM and OMV values [in base 10], as well as their general logarithmic status. Clearly there are no hard and fast rules here, except the observation that non-logarithmic data is typically associated with very low values of QTM. A glaring and embarrassing exception is  $k/x$  over (1, 1000) which has the rather large 251.2 value for QTM, yet it is not logarithmic in the general sense [except very narrowly on Earth]. Figure 7.35 depicts one possible explanation of why sufficiently large range on the log-axis [order of magnitude and by extension QTM] is not a guarantee of logarithmic behavior, even when such a measure is calculated only after excluding 10% on the left and 10% on the right. Here, the relatively large dent in the center precludes logarithmic behavior, strongly misleading that zealous statistician who adheres to OMV or QTM measure as the sole criterion of logarithmic behavior, come what may. Such a possibility should be

Data Set	Q <sub>90%</sub> / Q <sub>10%</sub>	General Status	Earth's OMV
Time Between Earthquakes	28.2	Logarithmic	1.5
USA Population Centers	112.4	Logarithmic	2.1
Exoplanets Mass - 835 Discovered	201.6	Nearly Logarithmic	2.3
Pulsar Rotation - 2210 Frequencies	268.0	Nearly Logarithmic	2.4
Chemical Molar Mass - 2175 compounds	5.7	Nearly Logarithmic	0.8
Breaking a Rock 12 Times into 4096 Pieces	4751.6	Logarithmic	3.7
Planet Formation P=0.25 15000 → 3882	12.7	Logarithmic	1.1
All Possible (1296) Products of 4 Dice	22.5	Logarithmic	1.4
Price List - Canford Audio PLC UK	266.8	Logarithmic	2.4
Price List - www.TheEventLine.com	42.6	Nearly Logarithmic	1.6
Price List - http://www.mdhelicopters.com/	2060.0	Logarithmic	3.3
US County Area Data	6.5	Non-Logarithmic	0.8
LOG Symmetrical Triangular (1, 3, 5)	160.2	Logarithmic	2.2
Lognormal L = 5 S = 0.3	2.2	Non-Logarithmic	0.3
Lognormal L = 2 S = 0.6	4.7	Non-Logarithmic	0.7
Lognormal L = 8 S = 1.3	28.0	Logarithmic	1.4
Lognormal L = 0.8 S = 2.2	281.5	Logarithmic	2.4
Lognormal L = 3 S = 2.7	1029.0	Logarithmic	3.0
$k/x$ over (1, 10)	6.3	Non-Logarithmic	0.8
$k/x$ over (1, 1000)	251.2	Non-Logarithmic	2.4
$k/x$ over (0.000008, 22)	137863.9	Logarithmic	5.1
$k/x$ over (7, 517134877)	1971535.6	Logarithmic	6.3
Exponential Growth, B=1.5, F=1.01	3.7E+34	Logarithmic	34.6
Varied Data - Hill's Model	168592.4	Logarithmic	5.2
Chain U(U(U(U(U(0, 5666))))))	242.9	Logarithmic	2.4

Figure 7.34 Quantitative Order of Magnitude and Status for a Variety of Data Sets

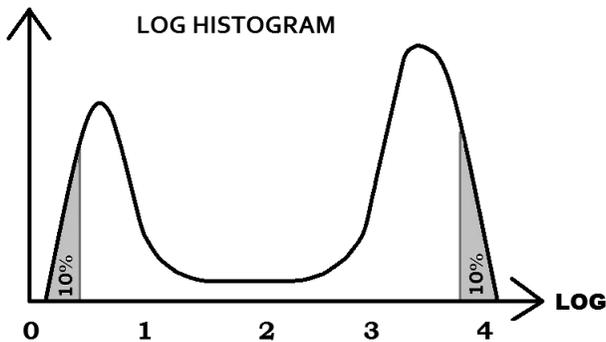


Figure 7.35 Deceptive Wide Range on the Log-Axis with Most of the Center Missing

compared with Fig. 4.43 in Section 4 regarding deceptive wide range on the log-axis when outliers and margins are included, which has led us to reformulate OOM as OMV by excluding the top 10% and the bottom 10% of data. Both figures represent two sharply opposing manifestation of the data. Figure 7.35 strongly emphasizes the edges and diminishes the center. Figure 4.43 strongly emphasizes the center and diminishes the edges. The common thread going through both figures is the fact that they do not utilize the full potential of the entire range on the x-axis stretching from the minimum value to the maximum value. Figure 4.43 does not utilize the edges much, and inflates the importance of the center. Figure 7.35 does not utilize the center much and inflates the importance of the edges. Surely, situations such as in Fig. 7.35 are quite rare, while those such as in Fig. 4.43 are much more frequently encountered.

Yet, in spite of the pessimistic prognosis given to QTM regarding its ability to predict logarithmic behavior, Chapter 68 *The Remarkable Malleability of Related Log Conjecture* may nonetheless point the way towards a different conclusion, given some restrictions such as continuity, smoothness, gradualism, and so forth, all in conjunction with a one-sided data density falling to the right.

**In spite of the fact that bin proportions for real-life data were deduced from the generic  $k/x$  distribution leading to the General Law, random logarithmic data sets do not possess directly that  $k/x$  property of being directly proportional to  $1/X$ ; there isn't exact halving in density whenever  $x$  is doubled (hence all logarithmic random data sets cannot be characterized by Allaart's sum invariance principle). Yet random logarithmic data sets relate to  $k/x$  indirectly by way**

of having corresponding overall fall in density (in the aggregate) as measured via bin schemes. All random logarithmic data sets show graduation and development in their fall as shall be discussed in Chapter 140 on bin development pattern, and on Earth their LOG density appears Normal-like or as an upside-down U-like shaped curve. Only the LOG density of the k/x distribution is uniform and steady throughout.

Base invariance principle can now be interpreted as the principle of the universality of the meaning of ‘logarithmic-ness’. That no matter what bin scheme is used, no matter what base is applied in digital Benford’s Law, classification of a given data set is a constant and universal, namely measuring-system-invariant, and that a change in base, F, or D, does not revolutionize data classification. Perhaps better terms for the two distinct data types are: ‘cooperative data set’ and ‘individualistic data set’, instead of the labels ‘logarithmic data set’ and ‘non-logarithmic data set’ commonly in use.

As a concrete demonstration of the universality of data classification, six distinct bin schemes shall be performed for the data set on time between 2012 earthquakes, as well as for U.S. county area data, as shown in Fig. 7.36 [bin width w is set at 0.009 and starting point S is set at the origin for all six schemes]. While the logarithmic data set of the time between earthquakes consistently obeys all of

	D	F	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F	Bin G	Bin H	Bin I
<b>The General Law</b>	2	5	68.3%	31.7%							
<b>Earthquake Time</b>	2	5	68.5%	31.5%							
U.S. County Area	2	5	79.3%	20.7%							
<b>The General Law</b>	4	9	50.0%	23.2%	15.3%	11.4%					
<b>Earthquake Time</b>	4	9	50.7%	22.2%	15.3%	11.7%					
U.S. County Area	4	9	64.5%	21.9%	7.0%	6.6%					
<b>The General Law</b>	6	6.1	34.0%	20.9%	15.1%	11.9%	9.8%	8.3%			
<b>Earthquake Time</b>	6	6.1	34.2%	20.6%	14.6%	11.9%	9.9%	8.8%			
U.S. County Area	6	6.1	26.3%	14.3%	19.3%	18.6%	13.5%	8.1%			
<b>The General Law</b>	7	3.8	25.2%	18.8%	15.0%	12.5%	10.7%	9.4%	8.3%		
<b>Earthquake Time</b>	7	3.8	25.8%	18.7%	14.6%	11.9%	10.8%	9.8%	8.4%		
U.S. County Area	7	3.8	18.7%	18.6%	16.9%	15.7%	11.9%	8.7%	9.5%		
<b>The General Law</b>	9	3	18.3%	15.2%	13.0%	11.4%	10.1%	9.1%	8.3%	7.6%	7.0%
<b>Earthquake Time</b>	9	3	18.5%	15.5%	12.9%	11.5%	9.9%	8.8%	8.5%	7.2%	7.2%
U.S. County Area	9	3	18.0%	12.4%	13.7%	10.5%	10.0%	12.4%	8.0%	8.2%	6.7%
<b>The General Law</b>	9	10	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%
<b>Earthquake Time</b>	9	10	29.6%	18.3%	13.5%	9.9%	7.6%	6.2%	5.4%	5.0%	4.4%
U.S. County Area	9	10	19.7%	10.1%	8.5%	13.6%	11.7%	13.8%	9.4%	6.7%	6.5%

Figure 7.36 Earthquake Complies with All Bin Laws — County Area Violates All

these six bin laws [the General Law], non-logarithmic data set pertaining to U.S. county area is a serial violator of the law of relative quantities.

An alternative data classification of logarithmic-ness, without relying on any fit into group behavior of other data sets, and without comparing overall relative quantities to the generic  $k/x$  case, is simply the observed near-steady bin results for all (small) values of bin width  $w$  in all logarithmic data set, and regardless of  $D$  and  $F$  bin structure. As seen in Figs. 7.29 and 7.30, and discussed at the end of Chapter 134, such sharp dichotomy could be used in the definition of the criteria of logarithmic-ness. To recap, logarithmic data is resistant to small changes in width  $w$  — having consistent bin behavior; while non-logarithmic data is highly unstable under even tiny variations in the value of width  $w$  — not having any consistent bin behavior.

## F = D + 1 BIN SYSTEMS ON REAL DATA

### YIELD $\text{LOG}_{\text{BASE}}(1+1/d)$

As one final confirmation of the general bin theory developed here, the above group of nine real-life logarithmic data sets and abstract distributions are examined under bin systems closely mimicking positional number system with  $F = D + 1$  [having a variety of values other than  $D = 9$  and  $F = 10$ ]. Results are then compared to the general digital logarithmic expression  $\text{LOG}_{\text{BASE}}(1+1/d)$ . As expected, strong agreements with digital Benford's Law in several such bin schemes gave the desired confirmation. In one concrete example, the bin scheme of  $D = 6$ ,  $F = 7$ ,  $S = 0$ ,  $w = 0.0003$  cast for all nine data sets gave average bin proportions that came out remarkably close to Benford's Law, as depicted in Fig. 7.37.

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F
Time Between Earthquakes	36.4%	21.0%	14.5%	11.2%	9.2%	7.6%
USA Population Centers	35.4%	20.9%	15.0%	11.5%	9.2%	7.9%
LOG Sym. Triangular (1, 3, 5)	35.1%	20.5%	15.0%	11.8%	9.8%	7.8%
k/x over (1, 1000000)	36.3%	20.4%	14.8%	11.4%	9.5%	7.7%
Exp. Growth, B=1.5, F=1.01	33.9%	21.3%	15.2%	11.8%	9.6%	8.1%
Lognormal, Loc=5, Shape=1	35.4%	20.5%	15.0%	11.8%	9.4%	7.9%
Lognormal, Loc=9.3, Shape=1.7	35.6%	20.8%	15.0%	11.6%	9.1%	7.8%
Varied Data - Hill's Model	37.0%	22.2%	13.8%	10.2%	9.0%	7.9%
Chain U(U(U(U(0, 5666))))	35.9%	21.1%	14.8%	11.1%	9.4%	7.6%
<b>Average of 9 Sets</b>	<b>35.7%</b>	<b>21.0%</b>	<b>14.8%</b>	<b>11.4%</b>	<b>9.4%</b>	<b>7.8%</b>
<b>BL: <math>\text{LOG}(1+1/d)/\text{LOG}(7)</math></b>	<b>35.6%</b>	<b>20.8%</b>	<b>14.8%</b>	<b>11.5%</b>	<b>9.4%</b>	<b>7.9%</b>
<b>Gen. R.Q. Law D=6 F=7</b>	<b>35.6%</b>	<b>20.8%</b>	<b>14.8%</b>	<b>11.5%</b>	<b>9.4%</b>	<b>7.9%</b>

Figure 7.37 Bin Scheme  $D = 6$ ,  $F = 7$ ,  $S = 0$ ,  $w = 0.0003$  Corresponds Strongly to Benford

## THE REMARKABLE MALLEABILITY AND UNIVERSALITY OF BIN SCHEMES

---

The elderly and kind mentor of the Tau Cetian statistician was a retired professor trained in the old pre-revolutionary numerical school who had written extensively about Zikuma's Law. Hardliner and staunchly numero-conservative, he was involved in numerous counter-revolutionary plots and conspiracies against the newly installed numberless dictatorship. Upon learning of the strange theories and artificial constructions invented by his favorite disciple, he immediately set about to visit the poor statistician in the guarded house. He was still hoping to persuade him to repudiate his entire work, and had intended to harshly reprimand him should he insist on continuing such misguided investigations. Fortunately for both of them, the subversive and furious cry of the old mentor was not overheard by the guards outside as he shouted: "How on Tau-Ceti-f can you expand digital segments along the x-axis in such a haphazard and chaotic way instead of the orderly  $D + 1$  way practiced for millennia?!" As events unfolded, this very exhortation turned out to give birth to the most disorderly and messy bin schemes ever invented. The ex-chief statistician, in a moment of supreme inspiration, agitated and provoked by the harsh phrases uttered by his beloved mentor, retorted in a spiteful and hallucinatory voice "perhaps we should vary inflation factor  $F$  itself as well within a single bin scheme, expanding segments by totally different inflation factors, and let us see if there would still be any consistent bin-pattern left standing there across all logarithmic data types!". As efforts to revive the fainted and shocked mentor were intensifying all around him, the statistician oblivious to all of this, set about calmly to empirically test all Clarikian, Tau Cetian, and Earthly data sets under a new 6-bin scheme with varying inflation factors vector  $F = \{2, 3, 4, 5, 6, 7, F_{N-1} + 1, \text{etc.}\}$ , starting with  $S$  at 0, and with an initial width of 0.03. The table in Fig. 7.38 depicts a remarkably steady result across all logarithmic data sets, constituting yet another bin law (but one that is by far more daring and bizarre). To emphasize that only logarithmic data types obey this newly proclaimed bin law, two additional non-logarithmic data sets are added at the bottom, U.S. county area and Payroll data

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F
Time Between Earthquakes	36.2%	21.9%	14.5%	11.2%	9.0%	7.2%
USA Population Centers	37.0%	20.2%	14.4%	11.1%	9.3%	8.0%
LOG Sym. Triangular (1, 3, 5)	36.4%	20.9%	14.3%	11.1%	9.3%	8.0%
k/x over (1, 1000000)	35.2%	19.4%	14.6%	12.2%	10.0%	8.6%
Exponential Growth, B=1.5, F=1.01	36.8%	20.5%	14.3%	11.2%	9.1%	8.1%
Lognormal, Loc=5, Shape=1	35.5%	20.8%	14.4%	11.5%	9.7%	8.0%
Lognormal, Loc=9.3, Shape=1.7	38.4%	20.4%	14.0%	10.8%	8.9%	7.4%
Varied Data - Hill's Model	34.7%	21.3%	15.9%	12.4%	8.9%	6.8%
Chain U(U(U(U(U(0, 5666))))))	33.6%	21.1%	15.1%	11.9%	9.7%	8.6%
(NON-Logarithmic) US County Area	21.3%	21.8%	22.6%	15.2%	11.6%	7.4%
(NON-Logarithmic) Payroll Data	21.9%	47.8%	17.5%	6.8%	3.2%	2.9%
(NON-Logarithmic) Normal(177, 40)	69.2%	0.4%	1.3%	4.0%	9.2%	15.8%
(NON-Logarithmic) Uniform(5, 78000)	21.1%	15.6%	15.8%	15.7%	15.7%	16.1%
(NON-Logarithmic) k/x over (1, 10)	37.0%	13.8%	12.2%	14.0%	12.4%	10.6%

Figure 7.38 6-Bin Scheme — Varying  $F = \{2, 3, 4, 5, \dots\}$  (Start = 0 Width = 0.03)

sets, serving as contrasts. These two data sets are well-known on Earth for their adamant refusal to digitally obey the law of Benford. These two odd data sets indeed are quite determined to refuse to obey **any** quantitative law postulated either digitally on Earth or bin-wise on Tau-Ceti-f. State of Oklahoma payroll data of the Department of Human Services for the first quarter of 2012 serves as the particular payroll data here, and it can be found on their website <https://data.ok.gov/Finance-and-Administration/State-of-Oklahoma-Payroll-Q1-2012/dq17-zvab>. Only those 2189 rows from the column ‘Amount’ pertaining to the Department of Human Services are considered. As an additional check, two well-known distributions are added, the Normal and the Uniform, infamous for their stubborn anti-Benford behavior. Also k/x defined over (1, 10) is added as a stark and shocking demonstration of its essentially non-logarithmic nature, in spite of its pretension to be universally so in all Earthly Benford’s Law digital schemes using 10 as the base.

Encouraged by this result, he was then wondering if the scheme has achieved its stability and consistency due to the predictability and orderly manner with which inflation factors were increasing, namely as in  $F_{N+1} = F_N + 1$ . Curious to discover if another law could be stated applying a totally arbitrary inflation vector having no intrinsic pattern whatsoever, he then set out to test all data sets and distributions under a 5-bin scheme, starting at 0, with an initial width 0.007, having an arbitrary and finite inflation factor vector  $F_i = \{2, 3, 4, 2, 5, 3, 6, 3, 5, 7, 4, 2, 3, 2, 7, 8, 9, 7, 3, 6\}$ . Although expansion along the x-axis is normally achieved by way of infinitely applying a fixed inflation factor, here merely the width of the last bin cycle is sufficiency large to enclose the entire range of each

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E
Time Between Earthquakes	38.7%	21.8%	15.8%	12.6%	11.1%
USA Population Centers	36.0%	22.7%	16.5%	13.6%	11.2%
LOG Symmetrical Triangular (1, 3, 5)	36.6%	22.7%	16.7%	13.3%	10.7%
k/x over (1, 1000000)	34.2%	22.7%	17.4%	13.3%	12.4%
Exponential Growth, B=1.5, F=1.01	35.3%	23.0%	16.9%	13.5%	11.3%
Lognormal, Location=5, Shape=1	33.6%	23.3%	17.7%	13.9%	11.4%
Lognormal, Location=9.3, Shape=1.7	35.5%	23.2%	17.0%	13.3%	10.9%
Varied Data - Hill's Model	35.6%	22.2%	16.8%	12.8%	12.6%
Chain U(U(U(U(U(0, 5666))))))	34.0%	22.7%	17.5%	13.8%	12.0%
<b>(NON-Logarithmic)</b> US County Area	38.5%	14.5%	17.2%	13.5%	16.3%
<b>(NON-Logarithmic)</b> Payroll Data	45.6%	6.6%	8.5%	18.3%	21.0%
<b>(NON-Logarithmic)</b> Normal( 177, 40)	39.2%	2.5%	8.4%	20.4%	29.5%
<b>(NON-Logarithmic)</b> Uniform(5, 78000)	21.0%	21.3%	20.6%	21.4%	15.7%
<b>(NON-Logarithmic)</b> k/x over (1, 10)	31.0%	22.5%	17.9%	15.0%	13.5%

Figure 7.39 5-Bin Scheme — Arbitrarily Varying F (Start = 0 Width = 0.007)

data set, since  $5 \cdot (0.007)^{2 \cdot 3 \cdot 4 \cdot 2 \cdot 5 \cdot 3 \cdot 6 \cdot 3 \cdot 5 \cdot 7 \cdot 4 \cdot 2 \cdot 3 \cdot 2 \cdot 7 \cdot 8 \cdot 9 \cdot 7 \cdot 3 \cdot 6} > \text{Max}(j)$  of any data set  $j$  here. Hence this very finite net cast over a ‘small’ portion of the x-axis is still quite wide and sufficient for our purposes. Figure 7.39 depicts the remarkably steady result across all logarithmic data sets for this bin scheme with arbitrarily chosen F factors. The significant deviations of U.S. county area and Payroll data sets, the Normal, and the Uniform, demonstrate again their staunch anti-logarithmic stand.

Further emboldened by this remarkable result, the statistician then set about performing an even more radical and outrageous bin scheme by using arbitrary fractional values for the inflation factors  $F_i$ ! Figure 7.40 depicts the results of a 6-bin scheme, starting at 0, with an initial width of 0.037, and utilizing the arbitrarily chosen set of  $F_i$  inflation fractional factors: {2.37, 3.08, 1.55, 4.17, 1.18, 2.35, 1.82, 5.07, 3.39, 2.04, 4.82, 7.07, 2.33, 6.67, 3.01, 1.67, 2.97, 3.33, 6.08, 2.25}. The near-steady proportions here strongly suggests that there is no need whatsoever to fix  $F_i$  as integers in order to observe a common rate of fall in histograms for all logarithmic data types.

As his initial shock and excitement slowly waned, self-doubts emerged in the mind of the statistician, who wondered whether the variations within each of the last three bin schemes in Figs. 7.38, 7.39, and 7.40, however small, are systematic and indicative of a lack of any law or regularity. His faith fully returned when he realized that even within the disciplines of Benford’s and Zikuma’s Laws variations

Data Set	Bin A	Bin B	Bin C	Bin D	Bin E	Bin F
Earthquakes	27.7%	19.5%	16.3%	13.5%	12.3%	10.7%
USA Population	28.3%	20.2%	16.6%	14.0%	11.1%	9.8%
Symmetrical Triangular	29.0%	20.0%	15.8%	13.2%	11.7%	10.3%
k/x over (1, 1000000)	26.7%	20.8%	16.6%	13.8%	12.4%	9.8%
Exponential Growth	27.1%	19.8%	16.0%	14.1%	12.4%	10.7%
Lognormal, L=5, S=1	27.6%	20.4%	16.3%	13.3%	11.7%	10.7%
Lognormal, L=9.3, S=1.7	31.5%	20.5%	15.8%	12.3%	10.7%	9.1%
Varied Data - Hill's Model	28.2%	19.3%	16.7%	13.7%	12.5%	9.6%
Chain 5 Uniforms	26.0%	20.4%	16.1%	14.3%	12.4%	10.8%
<b>(NON-Logarithmic)</b> US County Area	26.6%	25.0%	16.5%	15.0%	9.3%	7.6%
<b>(NON-Logarithmic)</b> Payroll Data	31.3%	10.8%	14.4%	15.1%	15.3%	13.1%
<b>(NON-Logarithmic)</b> Normal(177, 40)	23.9%	42.9%	24.8%	5.2%	1.5%	1.8%
<b>(NON-Logarithmic)</b> Uniform(5, 78000)	19.3%	16.0%	16.3%	15.7%	16.5%	16.2%
<b>(NON-Logarithmic)</b> k/x over (1, 10)	21.1%	26.7%	21.6%	11.3%	10.0%	9.3%

Figure 7.40 6-Bin Scheme — Fractional Arbitrarily Varying F (Start = 0 Width = 0.037)

Data Set	1	2	3	4	5	6	7	8	9
Time Between Earthquakes	29.9%	18.8%	13.5%	9.3%	7.5%	6.2%	5.8%	4.8%	4.2%
USA Population Centers	29.4%	18.1%	12.0%	9.5%	8.0%	7.0%	6.0%	5.3%	4.6%
LOG Symmetrical Triangular (1, 3, 5)	30.1%	17.7%	12.3%	10.0%	7.9%	6.8%	5.6%	5.0%	4.5%
k/x over (1, 1000000)	30.5%	17.5%	12.4%	9.9%	7.7%	6.7%	5.8%	4.9%	4.7%
Exponential Growth, B=1.5, F=1.01	30.3%	17.7%	12.4%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%
Lognormal, Location=5, Shape=1	31.4%	17.7%	12.2%	9.1%	7.7%	6.5%	5.8%	4.9%	4.7%
Lognormal, Location=9.3, Shape=1.7	29.7%	17.6%	12.4%	9.7%	8.0%	6.7%	6.0%	5.2%	4.6%
Varied Data - Hill's Model	28.8%	16.4%	12.4%	9.8%	8.3%	7.3%	6.1%	5.7%	5.3%
Chain U(U(U(U(U(0, 5666))))))	30.0%	16.9%	12.5%	9.6%	8.5%	6.8%	5.7%	5.4%	4.6%
<b>Digital Benford: LOG(1 + 1/d)</b>	<b>30.1%</b>	<b>17.6%</b>	<b>12.5%</b>	<b>9.7%</b>	<b>7.9%</b>	<b>6.7%</b>	<b>5.8%</b>	<b>5.1%</b>	<b>4.6%</b>
<b>(NON-Logarithmic)</b> US County Area	16.2%	10.0%	10.7%	15.8%	15.2%	10.4%	8.6%	7.1%	5.9%
<b>(NON-Logarithmic)</b> Payroll Data	12.9%	41.2%	25.4%	8.3%	5.2%	2.8%	1.3%	1.8%	1.0%
<b>(NON-Logarithmic)</b> Normal( 177, 40)	69.2%	28.0%	0.1%	0.1%	0.1%	0.2%	0.4%	0.8%	1.2%
<b>(NON-Logarithmic)</b> Uniform(5, 78000)	14.1%	14.1%	14.2%	14.3%	14.2%	14.5%	11.7%	1.4%	1.4%

Figure 7.41 Digital Results for all Data Sets — Benford's Law Base 10

of such small magnitudes are common and do not invalidate the laws in any way. The table he had arranged as seen in Fig. 7.41 regarding typical results from digital BL, and the low magnitudes of deviations there, put his mind to rest and convinced him that such mild variations are the norm in many similarly stated statistical laws. The data set k/x defined over (1, 10) is intentionally omitted here being that relative to this particular Benfordian bin system of base 10 it does indeed comply with the law, a fact which helps it pretend to be universally logarithmic.

To totally put the whole issue of errors, outliers, and variations to rest, he took another step of systematically calculating SSD value for each data set in both schemes of Figs. 7.40 and 7.41 (arbitrary fractional factors bin scheme and digital Benford’s Law). To do that though, he first had to proclaim a new fractional bin law for the particular scheme in Fig. 7.40 by simply using the empirical average across all nine data sets. That average came out to be {28.0%, 20.3%, 16.2%, 13.5%, 11.8%, 10.1%} and deviations were then measured against this proportion vector. The results are displayed in the tables of Figs. 7.42 and 7.43 where SSD values for each particular data set are shown within the context of each law.

Although satisfied with the overall message of compliance in Fig. 7.42, compared with that of non-compliance in Fig. 7.43, he nonetheless was a bit troubled by the fact that the grand average across all logarithmic data types for Benford’s Law of 1.5 was about half of the grand average for fractional factors bin law of 3.6, and that the two grand averages for non-logarithmic data sets had by far more pronounced deviations from Benford’s Law than from his newly minted fractional

Data Set	Benford's Law	Fractional Factors
Time Between Earthquakes	3.1	1.0
USA Population Centers	1.3	1.2
LOG Symmetrical Triangular (1, 3, 5)	0.2	1.4
k/x over (1, 1000000)	0.4	2.7
Exponential Growth, B=1.5, F=1.01	0.0	1.8
Lognormal, Location=5, Shape=1	2.3	0.6
Lognormal, Location=9.3, Shape=1.7	0.2	17.0
Varied Data - Hill's Model	4.8	1.5
Chain U(U(U(U(U(0, 5666))))))	0.9	5.4
<b>AVERAGE OF ALL SSD:</b>	<b>1.5</b>	<b>3.6</b>

Figure 7.42 SSD for Benford and Fractional Bin Scheme — Logarithmic Data Sets

Data Set	Benford's Law	Fractional Factors
(NON-Logarithmic) US County Area	371.8	41.6
(NON-Logarithmic) Payroll Data	1085.5	124.0
(NON-Logarithmic) Normal( 177, 40)	2044.9	859.0
(NON-Logarithmic) Uniform(5, 78000)	453.3	154.9
<b>AVERAGE OF ALL SSD:</b>	<b>988.9</b>	<b>294.9</b>

Figure 7.43 SSD for Benford and Fractional Bin Scheme — Non-Logarithmic Data

bin law. He was wondering whether the sharper contrast between logarithmic and non-logarithmic data types within Benford's Law compared with milder contrast within fractional factors bin scheme was an indication that the fractional bin law is somewhat weaker. As a result, he took one final step of directly visualizing and comparing the results of these two laws for all nine logarithmic data sets as shown in Figs. 7.44 and 7.45. Admittedly Benford's Law looks a bit more consistent and slightly more orderly. Nonetheless, fractional bin scheme is also just about as neat, except that Lognormal with Shape = 9.3 and Location = 1.7 positions itself ever so slightly differently than everybody else.

Still restless and seeking a visual result rivaling Benford's consistency, he was at long last completely satisfied upon viewing Fig. 7.46, which depicts the 4-bin scheme seen earlier with fixed  $F = 8$  from the origin, having an initial width 0.0008 (Scheme A in Fig. 7.13). The most satisfying part was to see how the General Law (theoretical proportion of  $k/x$  infinitely expanded) on the leftmost side closely agrees with the rest of the empirical data. That numerical result was gotten by simply substituting 4 for  $D$ , and 8 for  $F$ , in the closed form expression of the law of relative quantities.

Besides visualizations, he was able to put the two bin schemes of arbitrarily varying inflation factors on some 'theoretical' basis as a confirmation of sorts by considering the average value of these varying  $F_i$  inflation values, thus 'enabling'

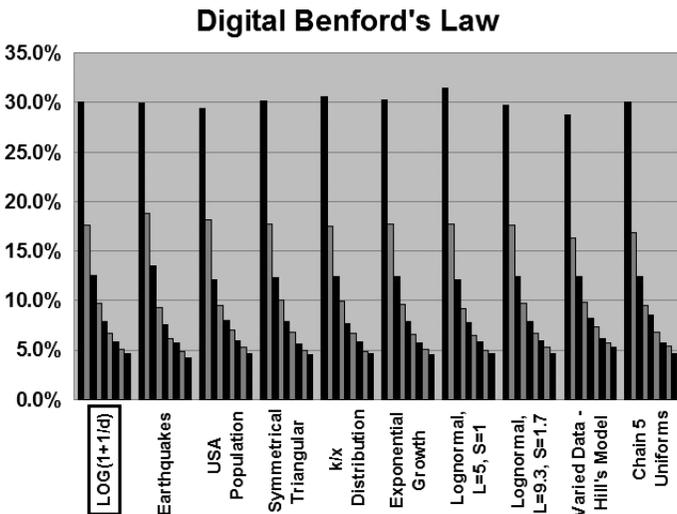


Figure 7.44 Visualizing Empirical Results from all Nine Data Sets — Benford's Law

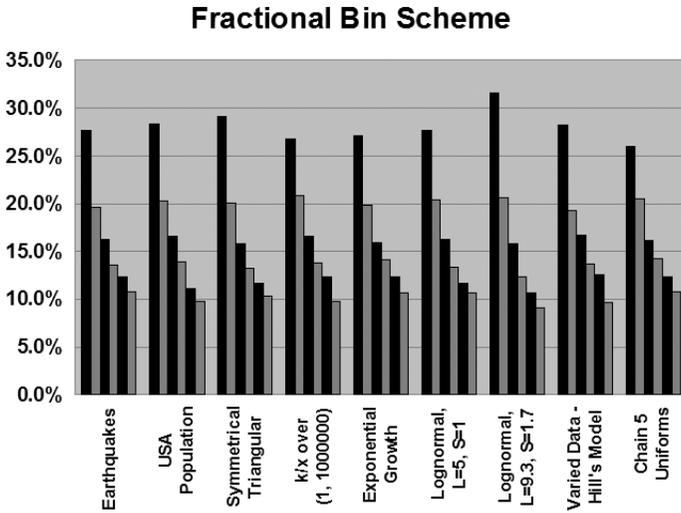


Figure 7.45 Visualizing Empirical Results from all Nine Data Sets — Fractional Bins

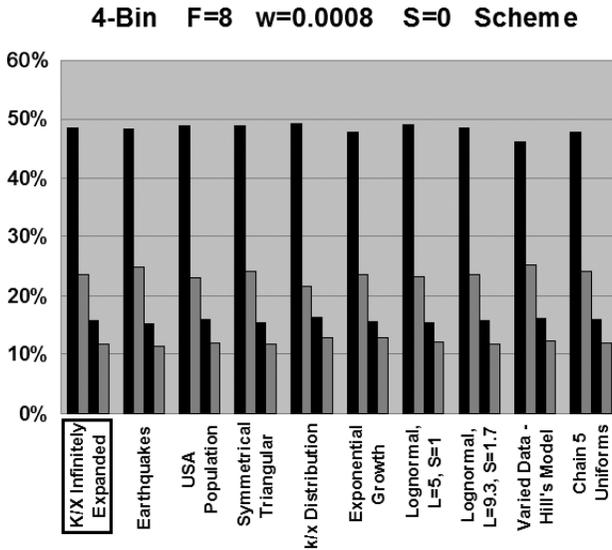


Figure 7.46 Visualizing Empirical Results from all Nine Data Sets — 4 Bins — F Fixed at 8

himself to insert this singular  $F_{AVG}$  value within the closed form expression of the General Law. Performing exactly this comparison (empirical to 'theoretical') for the two bin schemes, he obtained the following results:

Arbitrarily chosen set of integral  $F_i$  inflation factors are:

$$F_i = \{2, 3, 4, 2, 5, 3, 6, 3, 5, 7, 4, 2, 3, 2, 7, 8, 9, 7, 3, 6\}.$$

$F_{AVG}$  value for integral arbitrary factors is **4.55**.

Arbitrarily chosen set of fractional  $F_i$  inflation factors are:

$$F_i = \{2.37, 3.08, 1.55, 4.17, 1.18, 2.35, 1.82, 5.07, 3.39, 2.04, 4.82, 7.07, 2.33, 6.67, 3.01, 1.67, 2.97, 3.33, 6.08, 2.25\}.$$

$F_{AVG}$  value for fractional arbitrary factors is **3.36**.

Arbitrarily chosen integral  $F_i$  values (Fig. 7.39):

$$\text{Avg. of 9 sets empirical results: } \{35.6\%, 22.7\%, 17.0\%, 13.3\%, 11.4\%\}$$

$$\text{General Law } D = 5, F = \mathbf{4.55} : \{35.4\%, 22.9\%, 17.0\%, 13.5\%, 11.2\%\}$$

Arbitrarily chosen fractional  $F_i$  values (Fig. 7.40):

$$\text{Avg. of 9 sets empirical results: } \{28.0\%, 20.3\%, 16.2\%, 13.5\%, 11.8\%, 10.1\%\}$$

$$\text{General Law } D = 6, F = \mathbf{3.36} : \{27.4\%, 20.5\%, 16.4\%, 13.7\%, 11.7\%, 10.3\%\}$$

For both bin schemes, the 'theoretical' and empirical results are quite close, in spite of the fact that  $F_{AVG}$  value is just a simple average of the varying inflation factors and that its insertion in the expression of the General Law has no mathematical basis. This certainly lends arbitrarily varying  $F$  factors bin schemes some 'theoretical respectability'. Surely there exists no fractional base within a number system, nor could we conceive of a concept such as seven and a half bin scheme; both must be whole numbers, but fractional inflation factors are totally valid conceptual possibilities, even without the above agreement between empirical and 'theoretical' results. One must bear in mind that in the derivation of the series of algebraic expressions for the infinitely expanded  $k/x$  distribution no assumption was made restricting  $F$  to an integral value, therefore the whole mathematical edifice can be also employed for a scheme with a fixed fractional  $F$  value. Yet there is no mathematical basis whatsoever for employing and extrapolating the result of the infinite  $k/x$  model with steady  $F$  to a model with varying  $F_i$  (in spite of the apparent 'empirical' near correspondence in the numerical results above).

Great hopes sprang anew in the heart of the statistician. He wrote to the inquisition court advising of his recent undertakings and successful findings, arguing that his new bin schemes were so far removed from possible connection to any

number system whatsoever that a new trial was justified. But his arguments that such malleable schemes with arbitrary and varying fractional inflation factors are nothing but innocent constructions measuring relative quantities fell on deaf ears. The newly appointed inquisitor had to prove his credentials. To make things worse, the rebellion in the south by merchants and accountants clamoring for an easy way of counting and calculating and demanding a return to some kind of a reasonable number system threatened the already shaken revolution and precluded his release. Out of respect and consideration for his beloved and fragile old mentor, nothing about these outrageously new bin schemes was ever mentioned to him.

Unknown even to our daring statistician was the fact that even unequal bin schemes where some bins are consistently wider than others and some are narrower also lead to steady proportions and particular bin laws. In other words, that that sacred bin equality assumption may be violated and bin laws are still found in all logarithmic data sets! It is not necessary to empirically examine logarithmic data sets for such a pattern or to create any new mathematical edifice, since the following simple bin thought experiment (tale) would convince anyone of such straightforward consequence of the General Law. The tale relates to the awakening process of the bins themselves after eons of time of being in use on Tau-Ceti-f via the application of the very popular 5-bin and  $F = 2$  scheme, which grants the proportion vector of  $\{26.3\%, 22.2\%, 19.3\%, 17.0\%, 15.2\%\}$  to bins A, B, C, D, and E, respectively. The sudden and spontaneous gain in consciousness by the bins themselves immediately induces envy which consumes bin D and bin E who are jealous of the lower bins, and are especially resentful of bin A for attaining its high 26.3% proportion. Bins D and E then conspire against all the other bins and decide to merge in order to share their proportions, successfully gaining 32.2% share and attaining the new and desirable status of 'super bin DE'. Surely the General Law guarantees that proportion for super bin DE is 32.2% for all logarithmic data sets whenever DE is viewed as a singular wider bin. But all this is of course mere semantics and actual data portions falling onto bins is merge-invariant; the process of data allocation exists independently of any such bin naming and classification.

In another tale, the awakening bins are highly suspicious of each other and such lack of trust between bins precludes mergers and combinations. The most popular bin system on Tau-Ceti-f now is the 4-bin and  $F = 7$  scheme, which yields  $\{47.1\%, 24.2\%, 16.4\%, 12.4\%\}$  by the General Law. Bin A is consumed by greed and ambition. Its top proportion of 47.1% does not satisfy it in the least and it wishes to increase its high share even more. Knowing how morally weak and

corruptible Inflation Factor  $F$  happened to be, bin A then seduces it with a generous bribe in return for a promise for a preferential treatment in the form of a larger eightfold expansion on each cycle. Hence while all the other bins are expanding at the original rate of  $F = 7$ , our greedy bin A now expands faster by  $F = 8$ . Empirically, such a bin scheme yields highly consistent results for all logarithmic data sets examined [for any fixed value of initial width  $w$ ], and with very mild variation between them, allocating the vector of proportions of  $\approx \{73.0\%, 10.5\%, 8.8\%, 7.7\%\}$  to the great satisfaction of scheming and dishonest bin A.

In another very different tale, bin A feels guilty and apologetic for possessing the high 47.1% share granted to it by the General Law, and in a dramatic gesture of friendship has offered all the other bins self-reduction in its expansion rate, from 7 to 5. Hence while all the other bins are expanding at the original rate of  $F = 7$ , our friendly and modest bin A expands only by  $F = 5$ . Empirically, such a bin scheme yields highly consistent results for all logarithmic data sets examined [for any fixed value of initial width  $w$ ], with very mild variation, allocating the vector of proportions of  $\approx \{3.9\%, 53.6\%, 25.6\%, 16.9\%\}$ , and causing bin A to quickly retract and totally cancel its generous offer [it has not intended to give that much away to its peers]. Surely the whole mathematical edifice laboriously built here cannot be applied in this odd bin scheme where inflation factor varies within each cycle in favor of some bins while discriminating against other bins. A whole new mathematical calculation must be constructed again [supposedly] for the  $k/x$  distribution under such a scheme full of bin rivalry and intricate conspiracies.

Supporting the above empirical observation of a very steady bin proportions pattern for all logarithmic data sets where factor  $F$  favors some bins and discriminates against others, is another empirical observation testing non-logarithmic data sets here, and yielding very different results which are also highly unstable. For non-logarithmic data sets, proportions here vary wildly, being easily swayed by tiny changes in the value of the width  $w$  of the bins in the initial cycle, yet showing extreme stability whenever starting point  $S$  varies. Admittedly, for logarithmic data sets here as well, changes in  $w$  do affect results somewhat, especially when  $w$  nears the high width of 1, but such effects are not as dramatic as in the non-logarithmic case, and changes in  $S$  cause such minute and insignificant effects that they are barely noticeable, just as seen for non-logarithmic data. Such state of affairs regarding  $w$  and  $S$  with respect to logarithmic and non-logarithmic data sets, mirrors somewhat the state of affairs in normal equally inflated bins schemes, as summarized in

Figs. 7.29 and 7.30, except that bin laws here for logarithmic data must also incorporate an exact value for the initial width  $w$  in order to be properly stated.

It is not proper to extrapolate the above results of steady patterns seen in discriminating  $F$  inflation factors between bins, and incorporate arbitrary vectors of expansions constructed separately for each bin [i.e. fusing the idea of varying expansions rates between bins with the idea of arbitrary  $F$  factors]. This erroneous conjecture might tempt one to state that given a well-defined  $V_A = \{F_{A1}, F_{A2}, F_{A3}, \text{etc.}\}$  for bin A,  $V_B = \{F_{B1}, F_{B2}, F_{B3}, \text{etc.}\}$  for bin B,  $V_C, V_D$ , and so forth for all the bins, a steady and consistent bin proportions for all logarithmic data sets should then follow [i.e. bin laws]. This mistaken conjecture simply fails empirical examination performed on a variety of real-life logarithmic data sets and abstract distributions.

One should not lose sight of the fact that when bins are not made of equal width, bin laws may not convey any message about any possible fall in the data. Bins must start out with equal width between them, and keep maintaining that width-bin-equality during all relevant expansions, for us to be able to decipher any meaningful conclusion about fall in the density of the data and, by extension, about big versus small occurrences. Observing monotonically decreasing bin proportions does not necessarily imply that data is falling, but rather that perhaps bin A has expanded more than bin B, B more than C, C more than D, and so forth. If bin A expands more rapidly than bin B, then naturally it is more likely to capture more data falling onto it (in comparison with bin B). The ability of schemes with unequal bin sizes to convey a clear message about the data is confounded by the fact that they are irregular and uneven. Any message here is often more about the subjective measuring rod itself (i.e. the bin scheme) and less about the objective real phenomenon out there (i.e. data configuration).

## HIGHER-ORDER DIGITS INTERPRETED AS PARTICULAR BIN SCHEMES

The understanding of bin systems and their ability to measure fall in densities as developed here facilitates a radically different view about higher-order digit distributions in Benford's Law, allowing us to interpret them as some very particular arrangements of bin structures. For example, second-order digit distribution in base 10 as depicted in Fig. 7.47 can be viewed as the arrangement of equally spaced 10 inner bins within each outer 1st order bin — without any expansion until an IPOT point is encountered. Indirectly those secondary inner bins actually do expand, but only as a result of the fact that the main outer bin system (1st order) enclosing them expands after each IPOT encounter. In this arrangement, the outer cycles are situated between IPOT points of the first order, and within each bin of the outer cycle (1st order) we fully cycle one whole cycle of 10 inner bins (2nd order).

For example, within **each** outer bin [1, 2), [2, 3), [3, 4), ..., [8, 9), [9, 10), [10, 20), [20, 30), ..., [90, 100), [100, 200), [200, 300), and so forth, we cycle one full inner 10-bin structure, as seen in Fig. 7.47.

Applying such an interpretation, two explanations can be made of why second-order digit distribution in BL is more equal and less skewed than first-digit distribution: (I) The width of the inner bins of the second order is much smaller than those of the outer bins of the first order (on any segment of the x-axis), hence fall in density is registered as less severe, (II) Inner bins are not expanding (i.e. staying flat) until a new IPOT is reached, while outer bins are constantly expanding as they always span two adjacent IPOT points. The same reasoning can be applied



Figure 7.47 Second-Order Digits Viewed as Particular Inner Bins within Outer Bins

when third order is compared to second order, where width of third order is much smaller than width of second order, and third-order expansion is even less frequent than the frequency of the second order.

Alas, a much simpler view of the bin structure of digital second order is obtained when one forgets about outer bins and first digits altogether! When focusing only on the bin structure of the second order itself, one sees our very familiar bin structure but with a halting/vacillating process of expansion. Surely the bins stay flat for a while, but then they expand once, then stay flat for a while again, then expand once, and so forth, and all this is done in a well-structured and well-defined manner. A great deal of clarity is obtained when digital second-order distribution is presented simply in terms of a 10-bin scheme with varying inflation factors  $F_i$ . Referring to Fig. 7.47 again, and focusing only on the x-axis part over 1, the vector of  $F_i$  in a bin scheme tailor-made for second-order digits is as follow:

$$F_i = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \mathbf{10}, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \mathbf{10}, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \mathbf{10}, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \mathbf{10}, 1, 1, 1, 1, \text{etc.}\}.$$

For digital third order, those  $F_i = 10$  occurrences are literally fewer and far between, hence results show even more digital/bin equality. As our imprisoned Tau Cetian statistician already knows, any well-defined bin structure which alternates between flatness and expansion gives birth to a steady and consistent bin law, no matter how bizarre or disorderly  $F_i$  vector might appear. The ramification of such bin vista for higher-order digits in general is quite profound! It implies that any well-defined bin structure which alternates between flatness and expansion gives birth to a bin law such that **for a given (fixed) number of bins, resultant bin skewness depends on degree of flatness, namely a one-to-one relationship between skewness and degree of flatness. Flat bin schemes yield in general 1/D bin-equality (or some skewness in case of a prohibited very thick width but which is unstable as it depends on width value); fast-expanding bin schemes yield extreme bin inequality; while schemes with mild flatness having some expansion yield intermediate results.**

This crucial and fascinating result relates to one of the most central issues in Benford's Law and bin systems in general, namely that data falls on the x-axis in a multiplicative-like manner, just as seen in MCLT models. That there is an acceleration in how the spread of the data gets diluted towards the right side of high values, which cannot



<b>F Inflation</b>	<b>Bin A</b>	<b>Bin B</b>	<b>Bin C</b>	<b>Bin D</b>	<b>Bin E</b>	<b>Bin F</b>	<b>Bin G</b>
<b>1</b>	14.6%	14.2%	14.2%	13.9%	14.2%	14.2%	14.6%
<b>2</b>	19.2%	17.6%	14.6%	13.9%	12.1%	11.8%	10.7%
<b>3</b>	23.2%	17.7%	14.9%	12.7%	11.7%	10.5%	9.4%
<b>4</b>	25.5%	19.0%	15.1%	12.2%	10.4%	9.2%	8.5%
<b>5</b>	28.2%	19.2%	14.5%	12.0%	9.7%	9.0%	7.5%
<b>6</b>	30.0%	19.5%	14.5%	11.4%	9.2%	8.0%	7.3%
<b>7</b>	32.2%	19.2%	14.4%	11.2%	8.7%	7.7%	6.7%
<b>8</b>	33.4%	19.2%	13.6%	10.8%	9.0%	7.5%	6.6%
<b>9</b>	34.5%	19.4%	14.1%	10.5%	8.5%	6.9%	6.0%
<b>10</b>	35.6%	19.8%	13.2%	9.8%	8.5%	7.1%	6.0%
<b>11</b>	37.7%	18.9%	13.1%	10.1%	8.1%	6.3%	5.8%
<b>12</b>	38.1%	19.1%	13.1%	9.5%	7.8%	6.6%	5.8%

**Figure 7.48** Bin Skewness Increases in Direct Proportion to F — U.S. Population Data

decisively so) via bin schemes having a fixed F factor. In one such demonstration, several 7-bin schemes are performed on U.S. census population center data from the origin with initial width 0.0039, each with different F expansion factor. The results are shown in Fig. 7.48.

This decisive demonstration **cannot** be performed in the context of (*rigid*) Benford's Law where variations in F and D go hand in hand, since one is not allowed to vary independently of the other (they must relate to each other as in  $F = D + 1$ ). It is only in the context of (*flexible*) bin schemes that we can isolate F and exclusively vary it alone, observe this relationship and dependency, and thus demonstrate the principle.

**The last few chapters here clearly point to the conclusion that Base and Order in digital Benford's Law are mere variations on bin scheme structure! That is, Base and Order are incorporated into the very fabric of the bin system, constituting some of its parameters!**

The actual dependency of flat bin schemes ( $F = 1$ ) on the value of the width in real-life data [as opposed to our theoretical model build upon the ideal  $k/x$  distribution] shall be demonstrated by examining the census data on U.S. population centers. The table in Fig. 7.49 depicts a wide variety of results and skewness conditions of 4-bin flat schemes [all with  $F = 1$ ] depending on the value of the width. The starting point clearly does not play any role here unless it is placed very

Width	Starting	Bin A	Bin B	Bin C	Bin D
1	1	25.9%	25.2%	24.2%	24.7%
3	1	25.0%	25.6%	24.5%	24.9%
20	1	25.0%	25.3%	25.0%	24.7%
50	1	25.0%	26.0%	25.4%	23.7%
50	2	25.1%	26.1%	25.3%	23.6%
50	3	25.2%	25.8%	25.4%	23.6%
100	1	26.2%	26.9%	24.8%	22.1%
135	1	27.7%	27.8%	23.6%	20.8%
170	1	29.2%	28.1%	22.8%	20.0%
250	0	33.1%	27.6%	21.6%	17.7%
250	1	33.2%	27.5%	21.6%	17.7%
400	1	39.5%	26.6%	18.8%	15.1%
700	1	49.3%	23.4%	15.3%	12.0%
700	13	49.5%	23.3%	15.2%	11.9%
700	19	49.7%	23.1%	15.2%	12.0%
1200	1	58.8%	19.9%	12.5%	8.9%
2500	1	70.6%	14.8%	8.7%	6.0%
10 Million	1	100.0%	0.0%	0.0%	0.0%

**Figure 7.49** Results of Flat Bins ( $F = 1$ ) Depend on Width — U.S. Population Centers

far from the origin so that a significant portion of the data escapes the scrutiny of the applied bin scheme. For width size that is too small compared with the outlay of this population data, bin sizes are simply too refined to show any meaningful differentiation in data concentration, therefore discrete values fall upon the bins in a totally random manner, resulting in bin equality. For width size that is too big compared with the outlay of this population data, bin sizes are simply too crude, capturing significant portions of the data within the first bins of the first cycles, distorting their message of relative concentration. The huge width size of 10,000,000 shown at the bottom of the table of Fig. 7.49 captures the entire data within the first bin of the first cycle (allocating it 100%) and renders the whole bin scheme meaningless.

One must once again be reminded that there is no mathematical justification or basis for inserting a singular  $F_{AVG}$  value in the expression of the General Law for bin schemes with varying  $F_i$  values. For example, simple-mindedly extrapolating the mathematical results of  $k/x$  infinitely expanded with the singular insertion of  $F_{AVG}$  value for the bin system interpretation of the second-order distribution in



## BIN DEVELOPMENT PATTERN

---

---

For all **random** data types, a bin developmental pattern along bin cycles is certainly expected to emerge upon careful examination and regardless of expansion style F. Within the first cycle of bins on the left, approximate bin equality may prevail, or even a bin-reversal of fortunes. Around the central cycles where most of the data resides one should find bin proportions closely matching the relevant bin law given parameters D and F. Finally, around the far right region, more extreme bin inequality should prevail.

If one keeps in mind that skewness within any single cycle depends also on the relative width of the bin, and that thick bins show results that are more skewed than thin ones, then certainly this variation in skewness **within** a single bin scheme (bin development) is in perfect harmony and nicely consistent with the variation in skewness **between** different schemes having a variety of F inflation factors, as seen in Fig. 7.48 of the previous chapter. This is so since even fast-expanding scheme with high F factor achieves its widest bin sizes on the right after numerous applications of such F factor, while in the beginning on the left it must start with a very narrow width and is yet to apply sufficient number of such F multiplications.

Visualization of the above discussion is somewhat facilitated via Figs. 7.25 and 7.26, which show that too thick a net of bins is associated with over-skewness. Intuition dictates that, given a particular falling density curve, too thick and crude a net of bins cast upon the x-axis is associated with over-skewness as it records **macro fall**, while a refined and thin net is associated with less skewness as it records **micro fall**.

On the other hand, in the generic case of '**deterministic**'  $k/x$  defined over a very large range [and thus logarithmic], bin development depends on F expansion style. For a number-system-like expansion style with  $F = D + 1$ , bin proportions [skewness] are steady and consistent throughout the entire bin-cycle structure, with each bin cycle containing the same amount of overall data, just as was seen for our number system with base 10 in terms of digits and Benford's Law throughout the book. For a fast-expanding bin style where inflation factor  $F > D + 1$ , bin

proportions show development towards distributions that are more skewed, reaching a certain skewness level very rapidly, and then maintaining and continuing it thereafter in all subsequent cycles. Also data portion within each bin cycle increases but rapidly reaches a certain level, maintaining it about constant thereafter in all subsequent cycles. For a slow-expanding bin style where inflation factor  $F < D + 1$ , bin proportions show development towards less skewed and more even distributions (namely **an inverse development pattern!**) rapidly reaching a certain (milder) skewness level, and then maintaining and continuing it thereafter in all subsequent cycles. Also data portion within each bin cycle decreases, but rapidly reaches a certain level, then maintaining it about constant thereafter in all subsequent cycles.

All this is consistent with the theoretical expressions of the algebraic sequence of infinitely expanding bin cycles of  $k/x$  derived earlier. In fact it can be mathematically deduced from it. In the theoretical  $k/x$  case where  $F = D + 1$ , once- and twice-expanding schemes were already shown algebraically to reduce to the same bin proportions of  $\text{LOG}(1 + 1/d)/\text{LOG}(D + 1)$ , corresponding to non-expanding bin scheme. This reduction was assumed [inductively] to hold for any  $N$ -expanding scheme, an assumption vindicated by the successful reduction to a closed form and the short derivation of BL from the GL at the end of Chapter 131. All this implies constant bin proportions throughout all expanding cycles. This is so since a single expansion of  $k/x$  does not alter bin proportions of the original non-expanding  $x$ -axis range on the left, where value of  $k$  is admittedly reduced due to the expansion of the defined range [entire area must always sum to 1]. This reduction of the value of  $k$  (becoming  $k'$ ) affects height of  $k/x$  over original non-expanding range but not bin proportions there since area of each bin is reduced by an identical factor namely  $k'/k$ . Hence, if overall bin proportions stay unaffected with each expansion, it follows that with each expansion an extra set of identical bin proportions must be added (i.e. no bin development pattern). This is analogous to numbers being added one at a time where the average keeps its value as a constant, implying that all the numbers being added are identical (having the value of the average). Similar mathematical arguments can be made regarding  $F > D + 1$  scenario, showing that bin proportions there must be changing and becoming skewer with each expansion, as well as that milder and less skewed proportions are gotten with each expansion in the  $F < D + 1$  scenario.

For concrete examples of the above discussion, three different computer simulations of  $k/x$  realized data shall be performed for the three  $F$  scenarios above, with each scenario exactly fitting nine whole bin cycles. All realized  $k/x$  data starts at 3, and width of the bins in the first cycle is 3 as well, in all of the three scenarios (to ensure exact or better logarithmic behavior).

Cycle #	1	2	3	4	5	6	7	8	9
Bin A	43.2	43.4	43.4	42.1	43.2	45.0	42.4	43.7	42.7
Bin B	24.3	25.6	26.4	26.9	24.5	24.5	25.2	24.9	25.5
Bin C	18.1	17.7	17.9	17.4	18.4	17.8	17.8	19.0	17.7
Bin D	14.4	13.3	12.3	13.6	13.9	12.7	14.7	12.5	14.1
% of Data	11.0	11.1	11.0	11.4	11.0	11.1	11.3	10.9	11.3

Figure 7.50 Bin Development is Absent for Theoretical k/x distribution:  $F = D + 1$

**F = D + 1 scenario:**

$F = 5, D = 4, S = 3, w = 3$ , with nine bin cycles from 3 to 5859375.

Cycles are between: 3, 15, 75, 315, 1815, 9315, 46815, 234315, 1171815, 5859375.

20,000 values from the simulated k/x on (3, 5859375) are realized, providing the data set to be analyzed via bins. Results are shown in Fig. 7.50. Bin proportions (skewness) as well as percent of overall data within bin cycles are approximately constant and steady.

**F > D + 1 scenario:**

$F = 13, D = 3, S = 3, w = 3$ , with nine bin cycles from 3 to 7953374532. Cycles are between: 3, 12, 129, 1650, 21423, 278472, 3620109, 47061390, 611798043, 7953374532.

20,000 values from the simulated k/x on (3, 7953374532) are realized, providing the data set to be analyzed via bins. Results are shown in Fig. 7.51. Here, due to the rapidly expanding bin width, proportions are developing in the beginning with an increasing skewness on each successive cycle. In addition, percent of overall data increases in subsequent cycles. Although for both measures convergence to a steady level is quickly reached (roughly on the fourth cycle here).

Cycle #	1	2	3	4	5	6	7	8	9
Bin A	52.4	60.7	61.6	62.8	63.2	63.1	62.8	60.8	63.7
Bin B	28.5	24.2	24.9	23.5	22.6	23.3	23.2	23.3	22.3
Bin C	19.0	15.1	13.5	13.7	14.2	13.6	14.0	15.9	14.0
% of Data	6.3	10.9	11.6	11.9	11.4	11.8	12.2	11.9	12.2

Figure 7.51 Bin Development Pattern for Theoretical k/x distribution:  $F > D + 1$

**F < D + 1 scenario:**

$F = 2$ ,  $D = 8$ ,  $S = 3$ ,  $w = 3$ , with nine bin cycles from 3 to 12267.

Cycles are between: 3, 27, 75, 171, 363, 747, 1515, 3051, 6123, 12267.

20,000 values from the simulated  $k/x$  on (3, 12267) are realized, providing the data set to be analyzed via bins. Results are shown in Fig. 7.52. Here, due to the slowly expanding bin width, proportions show inverse development in the beginning (in the opposite direction of the norm for random data) with a reduction in skewness and towards more bin equality on each successive bin cycle. In addition, the percentage of overall data decreases. Although for both measures convergence to a steady level is quickly reached (roughly on the fifth cycle here).

In conclusion, for the case of the logarithmic distribution  $k/x$  defined over a very long range, bin development depends on the type of bin prism used in observing the fall of its density. Different bin prisms show different development patterns, depending on how fast bins are expanding, namely depending on the value of inflation factor  $F$  relative to the number of bins  $D$ .

In contrast with  $k/x$  distribution, all random data exhibit the same direction in bin development pattern (where skewness is constantly increasing), regardless of the type of bin prism used in viewing density fall [namely, for all  $F$  values relative to  $D$ ]. The tables in Figs. 7.53, 7.54, and 7.55 depict bin development patterns for the U.S. population centers data set for all the three scenarios regarding the value of  $F$  relative to the value of  $D$ . In all of these three scenarios (bin schemes), eight bins were used,  $F$  factors were chosen as 3, 9, and 15, the starting point  $S$  was set at 1, and the initial bin width  $w$  was chosen to be 2. For this population data set these choices should not be considered as too crude or large, since values increase

Cycle #	1	2	3	4	5	6	7	8	9
Bin A	30.9	20.7	18.3	19.0	16.6	15.0	15.6	18.1	17.3
Bin B	18.2	17.6	15.5	14.9	15.3	15.4	16.6	15.8	15.4
Bin C	13.7	13.1	13.7	13.8	14.0	14.8	14.0	13.6	13.6
Bin D	10.4	11.5	12.6	12.7	13.6	11.0	12.4	12.5	13.3
Bin E	8.3	10.8	10.5	12.4	9.7	12.7	11.6	10.7	11.3
Bin F	7.4	10.1	10.8	9.2	11.0	10.5	11.6	10.5	10.3
Bin G	5.8	8.1	9.6	9.9	10.4	10.6	9.2	10.7	10.0
Bin H	5.4	8.1	9.0	8.0	9.2	10.0	9.0	8.2	8.8
% of Data	26.8	12.3	9.6	9.1	9.0	8.4	8.5	7.8	8.5

Figure 7.52 Bin Inverse Development Pattern for Theoretical  $k/x$  Density:  $F < D + 1$

Cycle #	1	2	3	4	5	6	7	8	9	10
Bin A	10.7	10.4	11.9	17.7	22.6	24.8	23.7	24.8	32.1	31.3
Bin B	5.4	10.8	13.4	15.4	17.3	17.6	19.1	21.0	21.9	30.4
Bin C	10.7	9.0	12.4	14.1	14.3	14.5	14.0	15.0	14.1	9.8
Bin D	10.7	9.8	13.2	11.8	12.0	10.9	12.2	10.9	12.6	7.1
Bin E	21.4	11.7	12.8	11.7	9.9	9.4	9.4	9.6	7.2	6.3
Bin F	14.3	13.5	12.2	10.1	8.9	8.8	8.0	7.4	4.8	6.3
Bin G	10.7	16.2	11.1	10.1	7.7	7.4	6.7	5.5	4.4	3.6
Bin H	16.1	18.6	12.9	9.2	7.3	6.5	6.8	5.8	2.9	5.4
% of Data	0.3	2.5	11.6	22.9	23.8	17.5	11.8	6.3	2.7	0.6

Figure 7.53 Bin Development Pattern for U.S. Population Centers;  $F = 3$ ;  $D = 8$ ;  $F < D + 1$

Cycle #	1	2	3	4	5	6	7
Bin A	10.7	7.3	26.3	40.3	48.5	65.3	55.6
Bin B	5.4	8.5	18.9	18.8	20.1	12.9	22.2
Bin C	10.7	12.2	14.7	12.2	10.6	8.4	11.1
Bin D	10.7	15.1	11.9	8.7	7.6	4.5	0.0
Bin E	21.4	14.1	9.2	7.0	5.2	5.0	0.0
Bin F	14.3	14.9	7.5	5.1	3.4	3.0	0.0
Bin G	10.7	14.7	6.3	4.4	3.0	0.5	11.1
Bin H	16.1	13.1	5.3	3.5	1.7	0.5	0.0
% of Data	0.3	10.3	44.8	32.3	11.2	1.0	0.05

Figure 7.54 Bin Development Pattern for U.S. Population Centers;  $F = 9$ ;  $D = 8$ ;  $F = D + 1$

Cycle #	1	2	3	4	5	6
Bin A	10.7	7.3	39.3	52.9	76.4	72.7
Bin B	5.4	12.1	20.4	18.8	13.2	18.2
Bin C	10.7	13.9	12.1	9.6	3.3	0.0
Bin D	10.7	14.0	8.8	6.9	2.3	0.0
Bin E	21.4	13.8	6.8	4.3	1.4	9.1
Bin F	14.3	12.6	5.1	2.9	1.9	0.0
Bin G	10.7	13.6	4.2	2.3	0.5	0.0
Bin H	16.1	12.5	3.2	2.2	0.9	0.0
% of Data	0.3	17.7	54.8	24.2	2.9	0.1

Figure 7.55 Bin Development Pattern for U.S. Pop.;  $F = 15$ ;  $D = 8$ ;  $F > D + 1$

integrally and there is no data falling below 1 [namely, there exists no empty city or town, nor partial persons such as  $\frac{1}{4}$  of an inhabitant]. As can be seen clearly in these three tables, bin development pattern exists no matter how large an inflation factor  $F$  is chosen relative to  $D$  bin number. However, high inflation factors  $F$  relative to  $D$  are associated with faster and more dramatic bin development. Data congregation, however, always occurs around the central bins and very little data is found on the leftmost or rightmost bins, regardless of bin scheme style. This central congregation of random data is also sharply contrasted with  $k/x$  distribution which allocates steady portions to all bin cycles after the initial convergence during the first few cycles.

## THE GENERAL SCALE INVARIANCE PRINCIPLE

---



---

Pinkham's scale invariance principle, originally stated in the context of digits and Benford's Law, is a much more general principle extendable to all bin schemes, namely that the vector of proportions of **any** given logarithmic data set  $X$  examined via **any** given bin scheme is the same as the vector of proportions of  $K \cdot X$  examined via the same bin scheme where  $K$  is **any** positive real number.

Empirical examinations of bin results for a variety of re-scaling of U.S. population centers data set confirm the general status of the scale invariance principle. Scheme A is used, namely  $D = 4$   $F = 8$   $S = 0$   $W = 0.0008$ . The table in Fig. 7.56 shows an almost steady set of bin proportions with mild variation. In the special case where re-scaling factor equals  $F$  (the 'base' 8) or  $F^{\text{INTEGER}}$  (integral powers of the 'base', i.e.  $8^2, 8^3, 8^4$ , etc.) bin results are totally unchanged and are exactly as that of the bin proportions of the original data set. This is why the first four rows in Fig. 7.56 are identical. The fact that bin proportions are unaffected by re-scaling by integral powers of  $F$  in bin schemes, or that digits are unaffected by re-scaling by integral powers of ten in our number system, is straightforward and does not necessitate the application of the scale invariance principle. For such special re-scaling cases the entire data set is being transformed forward to the right by whole bin/digit cycles, and this has no effect whatsoever on resultant proportions.

Empirical examinations of bin results for a variety of re-scaling of realizations from  $k/x$  distribution defined over (14, 231316943) also confirm the general status of the scale invariance principle. Scheme A is used, namely  $D = 4$   $F = 8$   $S = 0$   $W = 0.0008$ . The table in Fig. 7.57 shows an almost steady set of bin proportions with mild variation.

Employing the closed form expression of the General Law for Scheme A by substituting 8 for  $F$  and 4 for  $D$  we obtain {48.6%, 23.7%, 15.8%, 11.9%}, which is quite consistent with the tables in Figs. 7.56 and 7.57.

The two empirical bin results above lend considerable theoretical respectability to the general bin theory developed here. These results show that the bin theory

Data Set	Bin A	Bin B	Bin C	Bin D
USA Population Centers	48.9%	23.1%	16.0%	12.0%
8 * USA Population Centers	48.9%	23.1%	16.0%	12.0%
8 <sup>2</sup> * USA Population Centers	48.9%	23.1%	16.0%	12.0%
8 <sup>3</sup> * USA Population Centers	48.9%	23.1%	16.0%	12.0%
2 * USA Population Centers	49.0%	23.7%	15.5%	11.8%
3 * USA Population Centers	48.6%	23.8%	15.7%	11.9%
5 * USA Population Centers	48.4%	24.0%	16.1%	11.5%
11 * USA Population Centers	48.9%	23.4%	15.3%	12.3%
6.351 * USA Population Centers	48.2%	23.6%	15.9%	12.3%
7.943 * USA Population Centers	48.7%	23.1%	16.2%	12.0%
0.387 * USA Population Centers	48.5%	23.8%	15.8%	11.9%
Law of Relative Quantities D=4 F=8	48.6%	23.7%	15.8%	11.9%

Figure 7.56 General Scale Invariance Principle Applied to U.S. Population Data

Data Set	Bin A	Bin B	Bin C	Bin D
k/x over (14, 231316943)	48.8%	23.6%	16.0%	11.5%
8 * k/x over (14, 231316943)	48.8%	23.6%	16.0%	11.5%
8 <sup>2</sup> * k/x over (14, 231316943)	48.8%	23.6%	16.0%	11.5%
8 <sup>3</sup> * k/x over (14, 231316943)	48.8%	23.6%	16.0%	11.5%
2 * k/x over (14, 231316943)	48.5%	23.6%	16.0%	11.9%
9 * k/x over (14, 231316943)	48.8%	23.6%	15.8%	11.8%
11 * k/x over (14, 231316943)	48.5%	23.7%	15.9%	11.9%
15 * k/x over (14, 231316943)	48.7%	23.7%	15.6%	11.9%
8.042 * k/x over (14, 231316943)	48.8%	23.6%	16.0%	11.6%
1.807 * k/x over (14, 231316943)	48.7%	23.7%	15.9%	11.7%
0.068 * k/x over (14, 231316943)	48.8%	23.5%	16.1%	11.6%
Law of Relative Quantities D=4 F=8	48.6%	23.7%	15.8%	11.9%

Figure 7.57 General Scale Invariance Principle Applied to k/x Over (14, 231316943)

is in perfect harmony with the essential feature of scale invariability found in digital Benford's Law.

Let us mathematically prove the general principle of scale invariance for the two most crucial and relevant distributions within the context of Benford's Law, namely for k/x distribution as well as for the Lognormal distribution. In both cases the Distribution Function Technique shall be applied using somewhat similar notations as in Freund's book "Mathematical Statistics", Sixth Edition, Chapter 7.3.

We shall assume that the data set  $X$  (actualizations of real values from the distribution) is being transformed (re-scaled) by the factor  $K$  into  $Y = K * X$ .

In the distribution case  $\text{pdf}(X) = k/x$  over  $(a, b)$ , the monotonic transformation  $Y(X) = K * X$  implies that  $X(Y) = Y/K$ , and  $dx/dy = 1/K$ , hence:

$$\text{pdf}(Y) = \text{pdf}(X(Y)) * |dx/dy|$$

$$\text{pdf}(Y) = k/[Y/K] * |1/K| = k/Y \text{ over } (aK, bK),$$

Namely, multiplying  $k/x$  type data set by any constant  $K$  yields the same density form as well as the same parameter  $k$ , although the range is different. Moreover, the change in the range should not adversely affect logarithmic behavior whatsoever. To demonstrate that we shall consider two contexts:

(I) A short  $(a, b)$  range in the context of digits and Benford's Law with base  $B$  where  $a$  and  $b$  span exactly an integral exponent difference of the base, namely  $\text{LOG}_B(b) - \text{LOG}_B(a) = \text{INTEGER}$ .

(II) The generic definition of logarithmic-ness for  $k/x$  defined over a truly huge range applied mostly in the context of the bin model but also in BL and digits for any base and spanning any type of interval.

In the first context, the newly defined range is still of an integral exponent difference, since  $\text{LOG}_B(Kb) - \text{LOG}_B(Ka) =$

$$\text{LOG}_B(K) + \text{LOG}_B(b) - \text{LOG}_B(K) - \text{LOG}_B(a) =$$

$\text{LOG}_B(b) - \text{LOG}_B(a)$ , which is also an INTEGER as was assumed for the original data set.

In the second context, the newly defined range is still as huge if  $K = 1$ , much longer if  $K > 1$ , but even when  $K < 1$  the true measure of 'length' is the quantitative order of magnitude defined earlier as  $\text{QTM} = Q_{90\%}/Q_{10\%}$  and which is unchanged, since  $\text{QTM}_{\text{RE-SCALED}} = KQ_{90\%}/KQ_{10\%} = Q_{90\%}/Q_{10\%} = \text{OTM}_{\text{ORIGINAL}}$ . The same argument is applied for OOM and OMV measures.

Let us turn our attention now to the case of the Lognormal distribution, namely  $\text{pdf}(X) = [1/(X\sigma\sqrt{2\pi})]*e^{-1/2 * [(\ln(X) - \mu)/\sigma]^2}$ . The monotonic transformation  $Y(X) = K * X$  implies that  $X(Y) = Y/K$ , and  $dx/dy = 1/K$ , hence:

$$\text{pdf}(Y) = \text{pdf}(X(Y)) * |dx/dy|$$

$$\text{pdf}(Y) = [1/((Y/K)\sigma\sqrt{2\pi})]*e^{-1/2 * [(\ln(Y/K) - \mu)/\sigma]^2} * |1/K|$$

The constant  $K$  on the left and on the right cancels out and we are left with:

$$\text{pdf}(Y) = [1/((Y)\sigma\sqrt{2\pi})]*e^{-1/2 * [(\ln(Y) - \ln(K) - \mu)/\sigma]^2}$$

This can be rewritten as:

$$\text{pdf}(Y) = [1/(Y\sigma\sqrt{2\pi})]*e \text{ to } [(-1/2)*[(\ln(Y) - (\ln(K) + \mu))/\sigma]^2]$$

Namely another Lognormal distribution of the same shape parameter  $\sigma$ , but of a different location parameter  $\ln(K) + \mu$ .

Since logarithmic behavior of the Lognormal is determined solely by the value of the shape parameter (being nearly logarithmic whenever shape is roughly larger than 1), the fact that the shape parameter is totally unaffected under re-scaling confirms the scale invariance principle for this crucial and very relevant statistical distribution.

As discussed in Chapter 136, logarithmic-ness is an absolute and universal property of any data set or distribution, irrespective of number system in use, base, or bin scheme chosen. Consequently, guaranteeing logarithmic behavior in the context of digits and Benford's Law for any re-scaled Lognormal distribution implies guaranteeing compliance with any bin scheme chosen, namely the general scale invariance principle for bin systems as well (for this case at least).

To prove the scale invariance principle in extreme generality (and not only in the cases of  $k/x$  and Lognormal), Related Log Conjecture discussed earlier in the book can be employed. Since re-scaling any data set or distribution involves the multiplication of each value by an identical  $K$  factor, it follows that the transformation that related log density undergoes due to re-scaling of data is simply a translation by  $\text{LOG}_B(K)$  to the right [for  $K > 1$ ]. Such translation alters neither the shape of related log density nor its range on the log-axis (two crucial factors in logarithmic behavior). The graph in Fig. 4.47 of Chapter 70 clearly demonstrates this principle.

A formal demonstration can be given by considering re-scaling **any** logarithmic data set  $X$  by the factor  $K$  as in:

$$x_1, x_2, x_3, \text{ etc.} \rightarrow K*x_1, K*x_2, K*x_3, \text{ etc.}$$

In log terms, such re-scaling results in the following transformation:

$$\text{LOG}_B(x_1), \text{LOG}_B(x_2), \text{LOG}_B(x_3), \text{ etc.} \rightarrow$$

$$\text{LOG}_B(K*x_1), \text{LOG}_B(K*x_2), \text{LOG}_B(K*x_3), \text{ etc.}$$

Expanding those log expressions we obtain the transformation:

$$\text{LOG}_B(x_1), \text{LOG}_B(x_2), \text{LOG}_B(x_3), \text{ etc.} \rightarrow$$

$$\text{LOG}_B(x_1) + \text{LOG}_B(K), \text{LOG}_B(x_2) + \text{LOG}_B(K), \text{LOG}_B(x_3) + \text{LOG}_B(K), \text{ etc.}$$

Namely the transformation of the whole curve of the density of related log by the simple translation of  $\text{LOG}_B(K)$  to the right, leaving the shape of the curve and its span on the log-axis unaltered.

Since logarithmic-ness is a universal property of any data set, it follows that guaranteeing logarithmic behavior in digital BL sense for any re-scaled logarithmic data set also guarantees its compliance with any bin scheme chosen, namely the general scale invariance principle as well.

This very general result is certainly consistent with what was shown earlier about  $k/x$  and Lognormal distributions using the Distribution Function Technique.

The location parameter of the Lognormal was shown to shift under re-scaling by  $[\mu] \rightarrow [\mu + \ln(K)]$ , hence, since  $\text{Lognormal}(\text{location}, \text{shape}) = e^{\text{Normal}(\text{location}, \text{shape})}$  it follows that re-scaling induces the transformation

$$e^{\text{Normal}(\mu, \sigma)} \rightarrow e^{\text{Normal}(\mu + \ln(K), \sigma)}$$

Converting to base B logarithm we obtain

$$\text{LOG}_B e^{\text{Normal}(\mu, \sigma)} \rightarrow \text{LOG}_B e^{\text{Normal}(\mu + \ln(K), \sigma)}$$

Using the log identity  $\text{LOG}_A(X) = \text{LOG}_B(X)/\text{LOG}_B(A)$  this is then expressed as:

$$\text{LOG}_e e^{\text{Normal}(\mu, \sigma)} / \text{LOG}_e(B) \rightarrow \text{LOG}_e e^{\text{Normal}(\mu + \ln(K), \sigma)} / \text{LOG}_e(B),$$

$$\text{or: Normal}(\mu, \sigma) / \text{LOG}_e(B) \rightarrow \text{Normal}(\mu + \ln(K), \sigma) / \text{LOG}_e(B),$$

namely a shift of related log base B by:

$$\ln(K) / \text{LOG}_e(B) = \text{LOG}_e(K) / \text{LOG}_e(B) = \text{LOG}_B(K) / \text{LOG}_B(B) = \text{LOG}_B(K).$$

This implies that the entire density of related logarithm base B of the Lognormal is shifted by  $\text{LOG}_B(K)$ , just as was shown in general earlier.

In the case of  $k/x$  distribution, it was shown that under re-scaling by  $K$ , the transformed density is  $\text{pdf}(Y) = k/Y$  over  $(aK, bK)$ , hence by Proposition I its related log density is also uniform (flat). The shift in the range  $(a, b) \rightarrow (aK, bK)$  implies a shift by  $\text{LOG}_B(K)$  to the right of the entire density of related logarithm base B, since:

$$(\text{LOG}_B(a), \text{LOG}_B(b)) \rightarrow (\text{LOG}_B(aK), \text{LOG}_B(bK))$$

$$(\text{LOG}_B(a), \text{LOG}_B(b)) \rightarrow (\text{LOG}_B(a) + \text{LOG}_B(K), \text{LOG}_B(b) + \text{LOG}_B(K)).$$

## PARADOXES EXPLAINED

---

Let us explore some of the paradoxes and counter-intuitive sentiments against digital Benford's Law orthodoxy, all of which are explained away and ironed out as the phenomenon is viewed through its quantitative prism.

Chapter 98 on singularities in exponential growth series describes a set of growth rates where digits are decisively non-logarithmic when viewed from the perspective of base 10 number system. Yet, except for that obsession with digital distribution, one wonders why a rebellious growth rate of, say, 93.070% should be considered any differently than, say, the digitally logarithmic 40.000%? Each rate depicts a sequence of quantities which explodes upwards at a constant rate, its (discrete) histogram being skewed to the right with a tail falling off in the same manner and rapidity as  $k/x$ . It is merely the digital aspect within base 10 number system that biases us to discriminate against 93.070% rate. It suffices to switch to almost any other base to reverse that negative digital view of 93.070%. The mere act of switching and re-aligning our superficial man-made digital marks, casting that digital net differently, is all that is needed so that the 93.070% exponential series would not get mangled and stuck by those very same marks (net) measuring it.

Do anomalous exponential growth series exist at all within bin systems where  $F \neq D + 1$ ? In other words, can an exponential growth series tramp upon the x-axis in a repetitive way, constantly favoring some bins at the expense of others? A moment's thought would convince anyone that this cannot be done when  $F \neq D + 1$ .

In the digital sense where  $F = D + 1$ , the ratios of the starting and ending points of each cycle is a constant, namely  $D + 1$ , and therefore anomalous series can exist because one could calibrate the factor by which the series grows in such a way as to match exactly the factor of the steady growth of the digital framework measuring it. For bin systems with  $F$  not equal  $D + 1$  on the other hand, the ratios of the right point of the rightmost bin to the left point of the leftmost bin for the various cycles are

$$[(D+1)w]/[w]$$

$$[(D+1)w+DFw]/[(D+1)w]$$

$$[(D+1)w+DFw+DFFw]/ [(D+1)w+DFw]$$

and so forth, and are clearly not constant unless  $F = D + 1$ . Therefore, calibrating a certain growth to perfectly fit one whole particular bin cycle would result in it becoming immediately uncoordinated from the bin framework on the next cycle. For example, in base 10, the ratios of the right point of the rightmost digital bin to the left point of the leftmost digital bin are:  $10/1 = 100/10 = 1000/100 = 10000/1000 = 10 = \text{constant}$ , thus enabling anomalous growth rates to exist by exactly adjusting the cumulative factor of a few steps forward to 10, copying the growth rate of the very system intending to measure it, and thereby tricking it.

It was this particular result that decidedly tipped the balance in favor of bin schemes and against digital Benford in the long drawn-out academic battle between Tau-Ceti-f and Earth. As a result of increasing concern within The Galactic Internal Revenue Service Department about undetected tax fraud utilizing sophisticated schemes involving highly calibrated anomalous exponential growth series in interest-bearing bank accounts on quasi-outlaw planets, it finally issued a formal declaration in support of  $F \neq D + 1$  bin schemes as its official forensic tool in all fraud detection legal matters.

Allaart's sum invariance characterization of Benford's Law was severely criticized, and it was demonstrated that it can only be valid for logarithmic data of the deterministic flavor relating to  $k/x$  distribution standing exactly between two IPOT points, adjacent or non-adjacent. As a decisive counter example to the false assertion generalizing his characterization to any type of data, it was argued to contrast two cases,  $k/x$  falling over (10, 100) of Fig. 4.82 and  $k/x$  falling over (20, 200) of Fig. 4.83. Yet again, if one lets go of that obsession with digits, Allaart's conceptual insight — if only postulated quantitatively — is valid in both cases, and for **any**  $k/x$  defined over **any** range for that matter. All that is required here is to treat each segment of the  $x$ -axis below the density curve equally and fairly, regardless of what digit is leading the numbers. Hence the segment (30, 32) is treated as equal as the segment (74, 76) in the sense that both are having exactly the width of two units, in spite of the superficial difference in having 3 or 7 as the first digit. Figure 7.16 is a testament to this superior quantitative Tau Certain vista of  $k/x$  over (20, 200). Comparing the average value of  $k/x$  within the generic sub-interval  $(K, K+E)$  **anywhere** inside its defined range, we obtain:

$$\text{Average} = \int x * f(x) \, dx \text{ [from } K \text{ to } K+E]$$

$$\text{Average} = \int x * [k/x] \, dx \text{ [from } K \text{ to } K+E]$$

$$\text{Average} = \int k * dx \text{ [from } K \text{ to } K+E]$$

$$\text{Average} = k * [(K+E) - (K)]$$

$$\text{Average} = k * [E]$$

Hence average (or sum conceptually) is a constant feature for any sub-interval with fixed width  $E$ , and this is true for **any**  $k/x$  on  $(a, b)$  regardless of exact  $a, b$  boundary values, even when their 'exponent difference' is fractional, not integral, within the framework of some invented and arbitrarily imposed positional number system with base  $B$ .

## DIGITS SERVING AS QUANTITIES IN BENFORD'S LAW

---

If Benford's Law is all about relative quantities, then one wonders how the story of this phenomenon could ever been told by Frank Benford by way of digits! On the face of it, for example,  $R = 85$  represents a quantity that is by far less than, say,  $P = 18796$ , and yet in the context of Benford's Law  $R$  belongs to a first-digit category higher than that of  $P$ .

In summarizing, the conceptual confusion arises from the fact that:

$$R = 85 \text{ and } P = 18796$$

$$\text{Quantity}(P) > \text{Quantity}(R)$$

$$1^{\text{st}} \text{ Digit}(P) < 1^{\text{st}} \text{ Digit}(R)$$

The answer is that the whole BL edifice is embedded within a well-designed structure where digits are counted and considered in a well-coordinated way along the x-axis, namely as a bin structure. Digit 7 does have a higher quantitative significance than digit 3, but only when both numbers/digits reside between the same pair of adjacent IPOT points such as, say, (10, 100), namely when both reside within the same bin cycle. For example, 78 has a higher quantitative value than 39 while also representing a higher first significant digit as both numbers/quantities reside within the same bin/digital cycle.

**We digitally start anew on each arrival of IPOT number in terms of quantitative considerations. Put another way, for numbers falling exclusively within, say, the sub-interval (10, 100), first digit does signify quantity! And for numbers falling exclusively within, say, the sub-interval (100, 1000), first digit does signify quantity! And so forth. Therefore, in the aggregate where all such IPOT sub-intervals are considered collectively, first-digit proportions convey a quantitative message as well, being the average local quantitative configuration throughout the entire data set.**

For the Roman civilization utilizing Roman Numerals two millennia ago, the above reasoning and Benford's Law surely do not apply. But now, our bin scheme perspective prods us to ask different questions than the ones asked in the earlier part of the book [Chapter 57] such as "Which Roman Numerals would predominate and which would be relatively rare in real data?", and "Could one find any concise and elegant expressions such as  $\text{LOG}(1+1/\text{symbol})$  when all real-life values are substituted with their Romanic symbols?" **Analyzing Roman Numerals proportions in real-life data sets does not convey any quantitative message whatsoever, even if one could arrive at a concise algebraic expression here! Hence such an investigation is not very interesting.**

As our obsession about invented digits and mere symbols wanes, we would like to ask instead "What are the best possible bin system choices of D and F that we creatures of the 21st century would recommend to them so that they would not struggle much with their cumbersome arithmetic and be able to easily measure relative quantities occurring throughout their Mediterranean civilization?"

## FRANK BENFORD'S PROPHETIC WORDS

---

---

Generalizing what Benford and Newcomb have discovered about numbers, digits, and mantissas, to relative quantities by way of the bin theory, it is quite remarkable to note Benford's eloquently written final philosophical words, all of which turned out to be the correct view of the whole phenomenon. It reads: *“As has been pointed out before, the theory of anomalous numbers is really the theory of phenomena and events, and the numbers but play the poor part of lifeless symbols for living things”*.

## FUTURE DIRECTION

---

---

The grand quest for future research would be to mathematically connect all of these diverse logarithmic processes shown in the book to resultant bin spread, deducing the general law of relative quantities directly from the setup and parameters of the generating processes themselves, and to accomplish that for any generic D and F bin scheme. The obvious relevant processes are: (I) random repeated multiplication processes as in Multiplicative Central Limit Theorem, (II) random combinations of random variables or data as in Hill's super distribution, (III) chains of random distributions, (IV) random rock breaking, (V) random throws of balls into boxes, (VI) random star and planet formation processes, (VII) random linear combinations, and others. Since the General Law was postulated in terms of  $k/x$  and its related exponential growth series, no further work is necessary in those cases.

Perhaps the chief source of fascination with Benford's Law emanates from the fact that so many distinct processes and real-life data sets that on the face of it have nothing in common with each other, all lead to the same exact digital and quantitative proportions! This is hard to accept! The field is seeking that grand unification theory that would once and for all point to that elusive 'intrinsically possessed' logarithmic-ness property in all of these processes and data sets. Although as Mother Nature has set things up perhaps, nothing more to logarithmic-ness can be found here other than data simply obeying Benford's Law and the General Law and falling off in the aggregate as sharply and in the same manner as that generic  $k/x$  distribution.

A related quest would be to evaluate bin areas via fragmented definite integrals in the generic Lognormal case as a function of location and shape parameters, and to show that its bin proportions correspond exactly to the General Law, regardless of the value of the location parameter. This result should be obtained in a limiting sense as shape parameter goes to infinity, or for all practical purposes whenever shape is sufficiently high and approximately larger than 1 to guarantee logarithmic behavior in the approximate.

## THE UNIVERSAL LAW OF RELATIVE QUANTITIES

---

---

Our general model measuring relative quantities via bin schemes that expand steadily by way of a constant  $F$  factor [via either  $F \neq D + 1$  or  $F = D + 1$ ], leads to the G.L.O.R.Q., which is totally divorced from digits and independent of any number system. Yet this result does not tell the whole story of relative quantities in sufficient generality, as seen earlier in Chapter 138 on ‘The Remarkable Malleability and Universality of Bin Schemes’, where steadily increasing inflation factors or even arbitrarily varying  $F$  factors are employed. Since it was (successfully) postulated that the generic pattern in how relative quantities are found in nature is such that (in the aggregate) the frequency of quantitative occurrences is inversely proportional to quantity, leading to the  $k/x$  distribution as the mathematically fitting density, any arbitrarily-constructed bin scheme should point to a new law by way of the evaluation of definite integrals constructed for the  $k/x$  distribution defined over an infinite range ( $w, +\infty$ ). For example, for a bin scheme with a vector of expansion  $F_i = \{1 + G/1, 1 + G/2, 1 + G/3, 1 + G/4, \text{ and so forth}\}$ ,  $G$  being any positive real number, the **universal** law of relative quantities (U.L.O.R.Q.) should be obtained mathematically by evaluating definite integrals for such a bin scheme constructed for  $k/x$  defined over an infinite range, leading to a different infinite algebraic sequence than the one in Chapter 129, and hopefully to a closed-form expression of that law if a limit of that sequence can be found. The above postulate about the generic pattern of relative quantities is nearly certain to be successfully applied once again to any bin scheme, and this conjecture is strongly supported by empirical confirmation of bin-behavior of  $k/x$  for the varying factors as in Figs 7.38, 7.39, and 7.40, where realized values from the simulated  $k/x$  defined over the relatively large range of (1, 1,000,000) [albeit not infinite] corresponded nicely to the bin proportions of all the other logarithmic data sets. Several other empirical experimentations with odd and arbitrary bin schemes also yielded consistent results (i.e. bin laws) where  $k/x$  defined over the huge range of (3, 734,336,442) also nicely participated in the group logarithmic behavior.

Empirical applications of Bowley Skewness Measure on various logarithmic data sets did not yield a consistent value, although all measures came out decisively positive in the range between 0.3 and 1.0, and not a single logarithmic data set fell below 0.3. On the other hand, non-logarithmic data sets gave decisively lower values, in the range between 0 and 0.25 approximately. Bowley Skewness is defined as:

$$\text{Bowley Skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$$

Where  $Q_1$  denotes the 1st quartile [*A.K.A. the 25th percentile*], namely the value where the first (lowest) 25% of ordered data lies to the left of it,  $Q_2$  being the median where 50% of data falls to the left of it, and  $Q_3$  being the 3rd quartile [*A.K.A. the 75th percentile*] where 75% of data is found to the left of it. Of note here is the very high Bowley measure of near 1.0 gotten for all logarithmic exponential growth series as well as for data simulated from  $k/x$  distributions defined over sufficiently long ranges. Yet for the highly logarithmic time between 2012 earthquakes data set the value of 0.32 was gotten which appears surprisingly low; while for the U.S. population centers data the higher value of 0.64 was gotten. Apparently, Bowley skewness is too crude a measure, being that it attempts to condense the entire complex behavior of the data into a singular value, an impossible task. The G.L.O.R.Q. or the U.L.O.R.Q. on the other hand are more refined, being that they offer us a sophisticated long vector of  $D$  length to measure the related concept of relative quantities and the implied overall fall or rise in data histogram.

## DIALOGUE CONCERNING THE TWO CHIEF STATISTICAL SYSTEMS

---

As the immense effort by the Federation to reduce biological, religious, and planetary conflicts was bearing fruits, mathematics, science, technology, and tolerance spread throughout the Milky Way. Alas, this momentary galactic peace and highly flourishing scientific advancement and collaboration rapidly gave way to the worst inter-planetary quarrels and menacing new clashes over scientific and mathematical schools of thoughts. Academic and planetary pride seemed to induce more threatening conflicts than territorial ambitions or competing ideologies concerning economic systems. The gravest quarrel was a brewing dispute between Tau-Ceti-f and its numerous small planetary allies advocating the bin system on one side and, on the other, Earth with its fewer but more potent and technologically superior collaborators insisting on the primacy of number systems and digital proportions. As academic zeal reached feverish pitch levels and imminent war loomed, the Federation decided to conduct a galaxy-wide televised debate hoping that such open discussion would air differences and reduce the tension. The notable statisticians, Numericio from Earth and Quanticio from Tau-Ceti-f, were chosen to represent the two warring planetary factions.

**Numericio:** Dear Milkiwayans, Earth and all her loyal allies greet all of you and wish to convey their sincere and peaceful intentions. The efficiency of positional number systems is well-known throughout the galaxy. Earth is immensely proud to be among the first 1000 or so planets to first discover it. How else could you teach a child to multiply 358 by 2317 for example? Digits are of the essence and measuring the proportions of their occurrences in most data sets shows a great deal of regularity and order. Esteemed Quanticio, why not conveniently apply the number system in use with statistical laws and patterns?

**Quanticio:** Firstly, Tau-Ceti-f and all her friendly and humble planetary associates sincerely regret this galactic-wide friction caused by such minor academic issue, and

it is our hope that this debate would pave the way towards greater understanding and cooperation. And now about the issue itself, esteemed Numericio, could you describe the personal motivation of that famed yet unknown Arabic or Indian digit 0 inventor, the earthling who significantly propelled you towards a complete positional number system?

**Numericio:** We know very little about his or her personality, life, or even the nationality. The motivation was to enable us to skip over empty powers of the base in a written number, to the next meaningful quantity relating to another higher power. This idea was immensely important to us and all our subsequent development. For example, without positional number system Kepler would not have been able to deal arithmetically with Brahe's vast astronomical data and discover the three planetary laws. These laws in turn were probably essential for young Newton in discovering classical mechanics, and so forth.

**Quanticio:** So the 0 digit inventor was only interested in the efficiency of your number system?

**Numericio:** Yes, nothing else, of course. Why is this of any importance?

**Quanticio:** Did the person ever pause for a moment to contemplate the enormous significance of his or her invention? I mean the resultant future proportion of this digit 0 in real data, if his or her idea was to become the standard and eventually adopted? Like the 0 proportion in data on the number of people infected per town by the Bubonic Plague? Or, say, in Marco Polo's revenue streams, the greed that caused your Black Death in the first place by bringing ships and rats from afar.

**Numericio:** I admit, surely, future digital proportion was not on his or her mind at all. We were at that epoch hard-pressed to find any reasonable number system, to be able to calculate and count properly. We were in a hurry and thinking only along short term lines. The long-term implications of such number system and the resultant digital proportions in future use for data sets never entered our minds. Besides, statistical theory and data analysis was at its infancy back then, or rather not even conceived of as yet.

**Quanticio:** So then you admit that your number system was not originally designed to measure occurrences of relative quantities in the real world or any other statistical pattern, including the pattern of its own digital symbols!?

**Numericio:** Yes, I admit, of course our number system was not designed for that purpose at all. At least, not originally.

**Quanticio:** Then as I understand it, many centuries later, two of your scientists accidentally discovered the (same) digital pattern. The coincidence in both cases was an old logarithm book and the differentiated use according to digits.

**Numericio:** Yes, the scientific disciplines of Newcomb and Benford never prompted them to directly investigate big vs. small proportions. Such issues of relative quantities in large data sets never occurred to Earth as being interesting or significant. The focus was on causality, prediction, and especially technological applications. You might characterize Earth's attitude as eccentric or egoistic, but we are by far more interested in finding out, for example, how to physically manipulate our orbit to double or triple its size so as to avoid the imminent heating up and rapid expansion of our star, much more so than, say, to provide the Federation Census Headquarter with the exact value of our orbit so that it may perhaps fit nicely within the database of millions of other planets into some irrelevant quantitative law. One less planet in the database is not going to affect result whatsoever in any case, and besides, our esteemed generals suspect that such gathering of planetary data by the Federation is all about preparations for the next galactic war and have classified such data as a military secret. We are practical beings and even now most of the interest we have in this digital pattern is focused on discovering fraud since greed is becoming increasingly more pervasive as of late.

**Quanticio:** You have only yourselves to blame since you created the economic and numerical system that induces it. But what is really striking here is your haphazard and coincidental manner of discovering scientific truths, instead of the organized and focused scientific method used in our enlightened planet. A case in point is how haphazardly and only coincidentally you have discovered the theoretical foundation of the speed of light by way of almost random mathematical manipulations of the four equations describing electromagnetic phenomena, instead of directly deducing such speed from the understanding of the electromagnetic force itself and how it interacts with the fabric of space-time.

**Numericio:** Mediator, this is insulting! James Clerk Maxwell is one of our most respected and beloved scientist who has contributed enormously to our development. Our entire mode of communication is based solely on his discovery. Describing his work as accidental and haphazard is totally unacceptable to Earth and all her proud allies!

**Galactic mediator:** Esteemed Quanticio, please stick to the facts. The purpose of this debate is to induce cooperation and goodwill among member planets, not

more antagonism. The Federation would not allow any side to use this platform as a propaganda tool. Both of you should avoid mentioning political and social systems!

**Quanticio:** Fine, so then Numericio, to summarize your general attitude, Earth is not really interested in measuring **directly** relative occurrences of quantities in physical data sets, but rather she is content to simply observe the relative proportions of the occurrences of her digital symbols.

**Numericio:** I admit that we feel quite attached to our number system and its digits, but if the general pattern of relative quantities also manifests itself in the case of our digital proportions, what's wrong with sticking exclusively to this aspect of the law? Can't one be inferred from the other?

**Quanticio:** No! Your digit law can be directly inferred from our general law of relative quantities, but not vice versa!

**Numericio:** I agree, fine, but for bin systems with  $F = D + 1$  your planet and my planet obtain identical results. Isn't your argument all about mere semantics?

**Quanticio:** Not at all! Digits and bins are essentially two different concepts. Let me make an analogy. After a long drought, when the rain finally arrives, the thirsty people are all out there with plastic, metal and wooden containers to capture every drop of water. All of a sudden, an eccentric earthling arrives on the scene carrying an expensive and heavy safe made of hardened steel, keys in the six-digit secret code, opens the door, and places it there laying on its back to capture water. Wouldn't that strike you as rather odd?!

**Numericio:** How does this in any way relate to digital and quantitative laws?

**Quanticio:** Well, using digits — originally designed and invented in your number system to express numbers and ease calculations — to express relative occurrences of quantities, instead of measuring it directly by way of generic bins, is akin to the use of a safe to collect raindrops. The generic concept of rain collection is the application of containers, while the generic concept of the law of relative quantities is the application of abstract bins to collect occurring quantities. Surely safes could hold water just as do normal containers if you place them on their backs, and that's when the two distinct concepts of securing valuables and collecting raindrops temporarily intersect. Surely your peculiar digits could show one type of the true pattern of relative quantities, and that's when the two distinct concepts of utilizing

numerical symbols and measuring relative quantities temporarily intersect. Yet, in both cases the [intersecting safe or digits] approach is not really what one would naturally select. You do restrict yourself unnecessarily by refusing to investigate  $F \neq D + 1$  bin schemes which enrich and deepen the understanding of the law of occurrences of relative quantities and demonstrate its versatility and many of its aspects. Moreover, your total lack of awareness of the Universal Law of Relative Quantities implies that you are failing in even basic understanding of the whole phenomenon of relative quantities and its much wider applications. And besides, let me remind you that there is no need for any safes, locks, and keys in our peaceful planet, since there are no thieves in post-revolutionary Tau-Ceti-f, while you on Earth always have to struggle with theft and corruption, and contain greed and all its horrific consequences.

**Numericio:** I do confess, we are trained in the constant use of our own man-made numbers and digits all our lives, and so we erroneously tend to associate them directly with physical phenomenon. We tend to lose sight of the fact that it was us who invented them, not Nature. Yet esteemed Quanticio, your bloc should acknowledge the fact that our bloc has been truly open-minded and flexible in the sense that our law has been stated in extreme generality by considering other bases. The algebraic expression  $\text{LOG}_{\text{BASE}}(1+1/d)$  encompasses all planets and civilizations using positional number system regardless of base. We have never narrowly-mindedly insisted on 10.

**Quanticio:** It is our view that such inclusiveness, however noble and commendable, has caused you to have a false sense of confidence, and actually inhibited you from further investigating the phenomenon in its much wider quantitative form. You were led to believe that you have covered all cases, all bases, and that there was nothing more to the phenomenon. You were applying a wide variety of heavy metal safes to collect water; safes of all sizes with distinct lock-combination systems, but never the generic efficient containers made of plastic and wood. In any case, I sense that differences in our positions have been considerably narrowed, and that the two blocs could come to some kind of compromised agreement.

**Numericio:** I agree, esteemed Quanticio.

**Quanticio:** On a more profound level esteemed Numericio, the most crucial difference between our systems is that you wrap all quantities with your arbitrarily-invented numerical digits and symbols, and insist on measuring the occurrences of

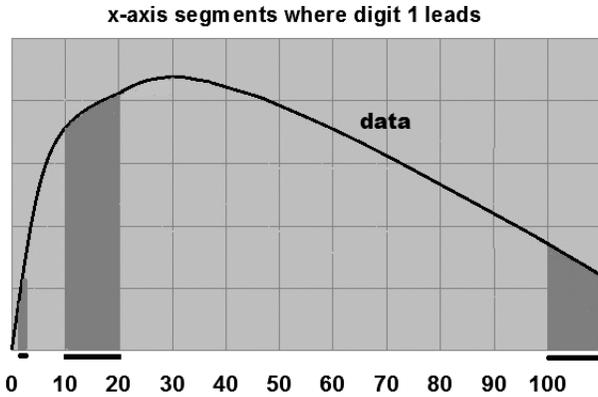
these envelopes, hence you are never convinced that the phenomenon has an independent existence; and in fact many misguided scholars on your planet still believe that Benford's Law is just a property of the number system itself (i.e. the envelop). We on the other hand are assured of its existence and have full faith in its independent reality, since we measure the naked quantities in their purest form without wrapping them around any invented symbols.

**Numericio:** I wholeheartedly accept your excellent argument esteemed Quanticio. In fact, ever since the news from your planet about the General Law has reached Earth, all these scholars have hurriedly abandoned their position and are now being discredited in their respective academic institutions.

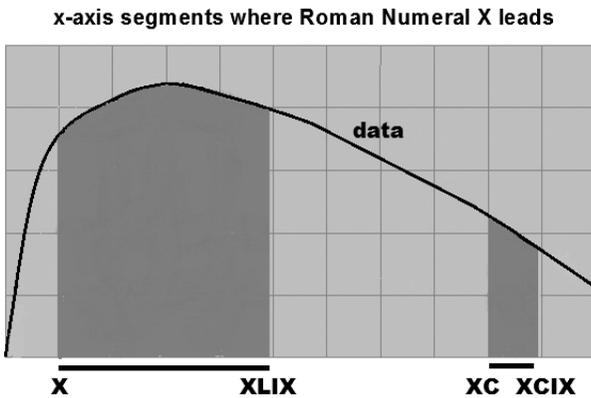
**Quanticio:** Let me make one final point about your earlier numerical system, namely Roman Numerals. Could any numeral pattern be found there?

**Numericio:** No, there exists no pattern there at all. Actually an eccentric Earthly statistician recently has converted large logarithmic modern data sets into Roman Numerals in an attempt to find a Benford's Law-like pattern in the proportions of the first placed numerals. The result was a decisive failure and caused a great deal of embarrassment to Earth. Each logarithmic data set yields unique vector of first-numeral proportions; nothing in common could be found across all such data sets, and so no law could be stated for Roman Numerals. Why should anyone expect to find a pattern in such an inefficient and obsolete number system?!

**Quanticio:** Your general view is misguided, esteemed Numericio. Statistical patterns in data sets have nothing to do with the efficiency of the number system itself. Consistent statistical patterns and number efficiency are two distinct concepts. The reason for the failure of that errant statistician is that the Roman Numeral number system divides the x-axis into disorderly compartments first-numeral-like, not fitting into any bin structure. Your modern positional number system just happened to divide the x-axis into more orderly compartments that nicely fit  $F = D+1$  bin schemes. Your 0 digit inventor never contemplated his or her resultant compartmentation on the x-axis. It is just a coincidence that it does fit nicely into bin schemes, enabling the law of relative quantities to be applied. Debate Exhibit # $\Lambda$  depicts part of the x-axis partitioning under your new positional number system along first-digit lines, focusing on digit 1. Debate Exhibit # $V$  depicts part of the x-axis partitioning under your old Roman Numeral system along first numeral lines, focusing on numeral X. Debate Exhibit # $\Delta$  depicts part



**DEBATE EXHIBIT #A** First-Digit x-axis Partitioning — Positional Numbers Base 10



**DEBATE EXHIBIT #V** First Numeral x-axis Partitioning — Roman Numerals

of the conversion table between the two number systems. In a deeper sense, your observation of a consistent digital pattern springs from two coincidences, the first being, that by chance your new number system happened to partition the x-axis according to the bin structure of the general law of relative quantities. The second coincidence is the keen observation of Newcomb and Benford about the peculiar conditions of old logarithm books.

I	1
II	2
III	3
IV	4
V	5
VI	6
VII	7
VIII	8
IX	9
X	10
XI	11
XII	12
XIII	13
XIV	14
XV	15
XVI	16
XVII	17
XVIII	18
XIX	19
XX	20
XXI	21
XXII	22
XXIII	23
XXIV	24
XXV	25
XXVI	26
XXVII	27
XXVIII	28
XXIX	29
XXX	30
XXXI	31

XXXII	32
XXXIII	33
XXXIV	34
XXXV	35
XXXVI	36
XXXVII	37
XXXVIII	38
XXXIX	39
XL	40
XLI	41
XLII	42
XLIII	43
XLIV	44
XLV	45
XLVI	46
XLVII	47
XLVIII	48
XLIX	49
L	50
LI	51
LII	52
LIII	53
LIV	54
LV	55
LVI	56
LVII	57
LVIII	58
LIX	59
LX	60
LXI	61
LXII	62

LXIII	63
LXIV	64
LXV	65
LXVI	66
LXVII	67
LXVIII	68
LXIX	69
LXX	70
LXXI	71
LXXII	72
LXXIII	73
LXXIV	74
LXXV	75
LXXVI	76
LXXVII	77
LXXVIII	78
LXXIX	79
LXXX	80
LXXXI	81
LXXXII	82
LXXXIII	83
LXXXIV	84
LXXXV	85
LXXXVI	86
LXXXVII	87
LXXXVIII	88
LXXXIX	89
XC	90
XCI	91
XCII	92
XCIII	93

XCIV	94
XCV	95
XCVI	96
XCVII	97
XCVIII	98
XCIX	99
C	100

DEBATE EXHIBIT #Δ Conversion Table — Roman Numerals and Base 10 Numbers

**This page intentionally left blank**

## REFERENCES

---

---

- Adhikari A., Sarkar B.** (1968). "Distribution of most significant digit in certain functions whose arguments are random variables". *Indian Journal of Statistics, Sankhya Series B*, 30, 47–58.
- Allaart, Pieter** (1997). "An Invariant Sum Characterization of Benford's Law". *Vrije Universiteit Amsterdam, Journal of Applied Probability*, 1997, 34, 288–291. <http://www.math.unt.edu/~allaart/papers/invar.pdf>
- Beber Bernd, Scacco Alexandra** (2012). "What the Numbers Say: A Digit-Based Test for Election Fraud". *Political Analysis* 20(2), 211–234. <http://pan.oxfordjournals.org/content/20/2/211>.
- Benford, Frank** (1938). "The Law of Anomalous Numbers". *Proceedings of the American Philosophical Society*, 78, 1938, 551.
- Breunig Christian, Goerres Achim** (2011). "Searching for Electoral Irregularities in an Established Democracy: Applying Benford's Law Tests to Bundestag Elections in Unified Germany". *Electoral Studies* 30(3) September 2011, 534–545.
- Buck Brian, Merchant A, Perez S** (1992). "An Illustration of Benford's First Digit Law Using Alpha Decay Half Lives". *European Journal of Physics*, 1993, 14, 59–63.
- Carslaw, Charles** (1988). "Anomalies in Income Numbers: Evidence of Goal Oriented Behavior". *The Accounting Review*, Apr. 1988, 321–327.
- Christian C.W., Gupta S.** (1993). "New Evidence on 'Secondary Evasion'". *The Journal of the American Taxation Association* 15, 72–93.
- Cleary Richard, Thibodeau Jay** (2005). "Applying Digital Analysis Using Benford's Law to Detect Fraud: The Dangers of Type I Errors". *Auditing: A Journal of Practice & Theory*. Volume 24, No. 1, May 2005, 77–81.
- Deckert Joseph, Myagkov Mikhail, Ordeshook Peter** (2011). "Benford's Law and the Detection of Election Fraud". *Political Analysis* 19(3), 245–268.
- Diaconis, Persi** (1977). "The Distribution of Leading Digits and Uniform Distribution Mod 1". *Annals of Probability* 5(1), 72–81. <http://statweb.stanford.edu/~cgates/PERSI/papers/digits.pdf>.
- Dorrell D. Darrell, Gadawski A. Gregory** (2012). "Financial Forensics Body of Knowledge". Wiley Finance.

- Durtschi Cindy, Hillison William, Pacini Carl** (2004). "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data". *Auditing: A Journal of Forensic Accounting*, 1524-5586/Vol. V( 2004), 17–34.
- Fewster, M. Rachel** (2009). "A Simple Explanation of Benford's Law". *The American Statistician*, Vol. 63, No. 1 (Feb 2009).
- Flehinger, J. Betty** (1963). "On the Probability that a Random Integer has Initial Digit A". *American Mathematical Monthly*, Vol. 73, No. 10 (Dec 1966), 1056–1061.
- Freund, John** (1999). "Mathematical Statistics". Sixth Edition. Pearson Education & Prentice Hall International, Inc.
- Gaines J. Brian, Cho K. Wendy** (2007). "Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance". *The American Statistician*, Vol. 61, No. 3. (2007), 218–223.
- Hamming, Richard** (1970). "On the Distribution of Numbers". *Bell System Technical Journal*, 49(8): 1609–25.
- Haw Mark** (2005). "Einstein's Random Walk". The School of Physics, University of Edinburgh, UK. <http://users-physics.au.dk/fogedby/statphysII/notes/Einstein's-random-walk.pdf>
- Henselmann Klaus, Scherr Elisabeth, Ditter Dominik** (2012). "Applying Benford's Law to individual financial reports. An empirical investigation on the basis of SEC XBRL filings". Working Papers in Accounting Valuation Auditing Nr. 2012–1. <http://www.econstor.eu/bitstream/10419/88418/1/773918388.pdf>.
- Hill, Theodore** (1995a). "Base Invariance Implies Benford's Law". *Proceedings of the American Mathematical Society*, 123: 887–95.
- Hill, Theodore** (1995b). "The significant-digit phenomenon". *The American Mathematical Monthly*, 102: 322–27.
- Hill, Theodore** (1995c). "A statistical derivation of the significant-digit law". *Statistical Science*, 10(4): 354–63.
- Hill, Theodore** (1998). "The First Digit Phenomena". *American Scientist*, Vol. 86, No. 4 (Jul–Aug 1998), 358(6).
- Hill, Theodore, Berger Arno** (2007). "Newton's Method Obeys Benford's Law". *The Mathematical Association of America, Monthly* 114 (Aug–Sept 2007), 588–601.
- Judge G. George, Lee Joanne, Cho K. Wendy** (2010). "Stigler's approach to recovering the distribution of first significant digits in natural data sets". *Statistics & Probability Letters*, Volume 80, Issue 2, 15 January 2010, 82–88.

- Kafri, Oded** (2009). "Entropy Principle in Direct Derivation of Benford's Law". (Mar 8, 2009), <http://arxiv.org/abs/0901.3047>.
- Kafri, Oded & Hava** (2013). "Entropy, God's Dice Game".
- Kossovsky, Alex Ely** (2006). "Towards A Better Understanding of the Leading Digits Phenomena". City University of New York, <http://arxiv.org/abs/math/0612627>.
- Kossovsky, Alex Ely** (2012). "Statistician's New Role as a Detective — Testing Data for Fraud". University of Costa Rica Publications — Ciencias Económicas 30-No.2: 2012/179-200/ISSN: 0252-9521. <http://revistas.ucr.ac.cr/index.php/economicas/article/view/8015/7634>.
- Kossovsky, Alex Ely** (2013). "On the Relative Quantities Occurring within Physical Datasets". (May 8, 2013), <http://arxiv.org/abs/1305.1893>.
- Leemis Lawrence, Schmeiser Bruce, Evans Diane** (2000). "Survival Distributions Satisfying Benford's Law". *The American Statistician*, Vol. 54, No. 4 (Nov 2000), 236–241.
- Leuenberger Christoph, Engel Hans-Andreas** (2003). "Benford's Law for the Exponential Random Variables". *Statistics and Probability Letters*, 2003, 63(4), 361–365.
- Ley Eduardo** (1996). "On the Peculiar Distribution of the U.S. Stock Indices Digits". *The American Statistician*, Volume 50, No. November 4, 1996, 311–313.
- Linville Mark** (2008). "The Problem of False Negative Results in the Use of Digit Analysis". *The Journal of Applied Business Research*, First Quarter 2008 Volume 24, Number 1.
- Logan Jonathan, Goudsmit Samuel** (1978). "The First Digits Phenomena". *Proceedings of the American philosophical Society*, Vol. 122, No. 4, Aug 1978, 193–197.
- Majola Gava Alexandre, Roque de Souza Vitiello Luiz** (2007). "Inflation, Quarterly Financial Statements and Fraud: Benford's Law and the Brazilian Case". [http://www.anpad.org.br/diversos/trabalhos/EnANPAD/enanpad\\_2007/CONT/2007\\_CONA819.pdf](http://www.anpad.org.br/diversos/trabalhos/EnANPAD/enanpad_2007/CONT/2007_CONA819.pdf).
- Mebane Walter** (2006). "Detecting Attempted Election Theft: Vote Counts, Voting Machines and Benford's Law". Paper prepared for the 2006 Annual Meeting of the Midwest Political Science Association, Chicago, IL.
- Mebane Walter** (2006). "Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections". *Proceedings of the Election Fraud Conference*, Salt Lake City, Utah, September 29–30, 2006.

- Miller, Steven** (2008). "Chains of Distributions, Hierarchical Bayesian Models and Benford's Law". (Jun 2008), <http://arxiv.org/abs/0805.4226>
- Newcomb, Simon** (1881). "Note on the Frequency of Use of the Different Digits in Natural Numbers". *American Journal of Mathematics*, 4, 1881, 39–40.
- Nigrini, Mark** (1992). "The Detection of Income Tax Evasion Through an Analysis of Digital Frequencies". Thesis, University of Cincinnati, Ohio.
- Nigrini, Mark** (1994). "Using Digital Frequencies to Detect Fraud". *The White Paper Index* 8(2), 3–6.
- Pinkham, Roger** (1961). "On the Distribution of First Significant Digits". *The Annals of Mathematical Statistics*, 1961, Vol. 32, No. 4, 1223–1230.
- Raimi, A. Ralph** (1969). "The Peculiar Distribution of First Digit". *Scientific America*, (Sep 1969), 109–115.
- Raimi, A. Ralph** (1976). "The First Digit Problem". *American Mathematical Monthly*, (Aug–Sep 1976).
- Raimi, A. Ralph** (1985). "The First Digit Phenomena Again". *Proceedings of the American Philosophical Society*, Vol. 129, No. 2, (Jun 1985), 211–219.
- Ross A. Kenneth** (2011). "Benford's Law, A Growth Industry". *The American Mathematical Monthly*, Vol. 118, No. 7, 571–583.
- Sambridge Malcolm** (2011). "Benford's Law of First Digits: From Mathematical Curiosity to Change Detector". *Sambridge Malcolm, Hrvoje T. and Pierre Arroucau, Asia Pacific Mathematics Newsletter*, (Oct 2011).
- Sambridge Malcolm** (2010). "Benford's Law in the Natural Sciences". *M. Sambridge, H. Tkali, and A. Jackson, Geophysical Research Letters*, Vol. 37, L22301, 2010.
- Saville, Adrian** (2006). "Using Benford's Law to Detect Data error and Fraud: An Examination of Companies Listed on the Johannesburg Stock Exchange". *Gordon Institute of Business Science, University of Pretoria, 2006. South African Journal of Economics and Management Sciences*, 9(3), 341–354. <http://repository.up.ac.za/handle/2263/3283>
- Shao Lijing, Ma Bo-Qiang** (2010a). "The Significant Digit Law in Statistical Physics". <http://arxiv.org/abs/1005.0660>, (May 6, 2010).
- Shao Lijing, Ma Bo-Qiang** (2010b). "Empirical Mantissa Distributions of Pulsars". <http://arxiv.org/abs/1005.1702>, (May 12, 2010). *Astroparticle Physics*, 33 (2010), 255–262.

- Stigler, J. George** (1946). “The Distribution of Leading Digits in Statistical Tables”. Written 1945–1946 and referred by Ralph Raimi as a non-published paper.
- Varian, Hal** (1972). “Benford’s Law”. *The American Statistician*, Vol. 26, No. 3.
- Weaver, Warren** (1963). “Lady Luck: The Theory of Probability”. Doubleday, Anchor Series, New York, 270–277.

**This page intentionally left blank**

## GLOSSARY OF FREQUENTLY USED ABBREVIATIONS

---

---

<b>LD</b>	Leading digits, or first significant digits
<b>BL</b>	Benford's Law
<b>FTD</b>	First-Two-Digits combination
<b>LTD</b>	Last-Two-Digits combination
<b>AGD</b>	Aggregate Global Data
<b>AGDI</b>	Aggregate Global Data Interpretation of Benford's Law
<b>SSD</b>	Sum Squares Deviation
<b>IPOT</b>	Integral Power Of Ten
<b>RLC</b>	Random Linear Combinations
<b>OOM</b>	Order of Magnitude — total units on the log scale — LOG[Max/Min]
<b>OMV</b>	Order of Magnitude of Variability — LOG(90th percentile/10th percentile)
<b>QTM</b>	Quantitative Order of Magnitude — 90th percentile/10th percentile
<b>CLT</b>	Central Limit Theorem
<b>MCLT</b>	Multiplicative Central Limit Theorem
<b>SGD</b>	Second-Generation Distributions
<b>LB</b>	Lower Bound
<b>UB</b>	Upper Bound
<b>LDIP</b>	Leading Digit Inflection Point
<b>PDF</b>	Probability Density Function
<b>RD</b>	Number of Relevant Digits for the particular chi-sqr test performed

**This page intentionally left blank**

# INDEX

---

- address data, 13, 14, 54, 210–215, 310–312
- Adhikari and Sarkar, 238
- Aggregate Global Data (AGD), 32, 210, 231, 234, 497–500
- Allaart, Pieter, 143, 146, 363–370, 583, 587, 620
- Andrews, George, 558, 565, 566
- anthropic principle, 387, 388
- arbitrarily varying F factors, 592, 593, 626
- artificial creation of logarithmic data, 507, 508
- Athena Guaranteed Futures, 100, 101, 136
- atomic weight, 380, 381
  
- balls & boxes, 402–414
- base (of a number system), 450, 451
- base invariance, 206, 230, 450, 451, 522, 523, 524, 575–581, 584, 588,
- Benford, Frank, 19, 20, 630, 634
  - attempted proof, 453, 471–473
  - exponential growth claim, 327
  - prophecy, 624
- Berger, Arno, 241
- bin development pattern, 608–613
- bin systems/schemes, 535–538
- bin-discriminating F factors, 599–601
- Bohemian Crystal Palace, 192–194
- Boltzmann, Ludwig, 377, 378
- Boltzmann-Gibbs distribution, 377, 378
- Bose-Einstein distribution, 378, 379
- Bowley skewness measure, 627
- Brahe, Tycho, 629
- Brownian motion, 378, 379
- Buck, Brian, 380
  
- Canford Audio PLC, 164–171
- carbon dioxide emissions, 141, 142, 281
- Carslaw, Charles, 79, 80, 97
- Cartesian coordinate system, 522, 523
- Central Limit Theorem (CLT), 65, 198, 389, 390
- chains — extrapolation of the second conjecture, 460
- chains of distributions, 219–226, 452–461, 462–470
  - as Benford’s own proof, 471–473
  - as explanation, 395–397
  - in balls & boxes model, 414
  - in hybrid causes, 419, 420
- chains — the infinite conjecture, 222, 455
- chains — the second conjecture, 224, 457
- chemical compounds, 381–385
- chemistry, 380–386
- chi-sqr test, 117, 119, 123–127
- Christian C.W., Gupta S., 80, 105
- Clarikia (planet), 532, 534, 544, 545
- Complex Averaging Schemes, 212–215
- computer expression for 1st digits, 506

- configurational entropy, 406, 407
- Corollary I, 262
- Corollary II, 263
  
- D — number of bins, 535
- Descartes, René, 522, 523
- Developmental Line, 152, 153, 168
- dice games, 13, 70, 71, 390–392
- Digital Development Pattern, 74, 75, 108, 109, 149–163, 176, 339–344, 345–348, 349–355, 356–359, 360–362, 494, 495, 503–505
- digit-anemic numbers, 110–112
- Dirac, Paul, 377, 378
- Dorrell and Gadawski, 140
- Durtschi *et al.*, 140
  
- earthquake, 35–40, 134, 135
- Einstein, Albert, 377–379
- election data, 94
- Engel, Hans-Andreas, 482
- entropy, 406, 407
- Excess Sum Digits 0 to 4 (ES04), 161, 168
- Excess Sum Digits 1 and 2 (ES12), 150, 167
- exoplanets, 34, 35, 282, 283, 418, 531
- Exponential Distribution, 183, 339–341, 346, 347, 480–483
  - chains of, 453–455, 482, 483
  - in nature, 377–379
  - product of, 394
- exponential growth series, 58–64, 67–69, 313–316, 317–326, 353–355, 363–366, 507, 508, 427–438, 619, 620
  
- F — inflation factor (bin schemes), 537
- factorial sequence, 241, 320
- Federal Reserve Bank (U.S.), 404, 405
  
- Fermi, Enrico, 377, 378
- Fermi-Dirac distribution, 378
- Fibonacci series, 239, 240, 320, 440, 441
- Financial Crimes Enforcement Network (FinCEN), 104
- financial statements, 89, 98, 99
- First-Three Digits, 25
- First-Two Digits, 23–26
- Flehinger infinitely iterated averaging scheme, 214, 218, 219, 453
- fractional arbitrarily varying F factors, 593, 594, 626
- Freund, John, 476, 615
  
- Gamma Distribution, 479
- General Form of Benford's Law, 256–258
- general law of relative quantities, 557, 561
- General Scale Invariance Principle, 614–618
- generalized Benford's Law for all bases, 450
- Gibbs, Josiah Willard, 377, 378
- Greek parable, 200–206, 310–312, 313–316, 529
  
- half-lives of radioactive materials, 380
- Hamming, Richard, 65, 66, 238, 239
- Haw, Mark, 378, 379
- higher order digits, 442–449, 602
- Hill super distribution, 226, 227–229, 275–277, 293, 294, 301, 304, 349, 389, 421, 422, 501–505
- Hill, Theodor, 32, 37, 80, 90, 227–229, 230, 234, 241, 275–278, 293, 294, 419, 447, 473, 501
- Hume, David, 530
- hybrid causes, 419, 420
- Hybrid Data, 197

- infinite algebraic sequence of  $k/x$  bin scheme, 554, 557
- Integral Power of Ten (IPOT), 57, 356–359
- Internal Revenue Service (IRS), 104, 105
- inverse bin development pattern, 609
- John Tukey's Biweight algorithm, 607
- Kafri, Oded, 406–413, 415
- Kepler, Johannes, 629
- Kossovsky, Alex Ely, 214, 219, 224, 323
- $k/x$  distribution, 246–252, 261–265, 305, 306, 307–312, 318–320, 579–581, 583–585, 614–618
  - applicability of, 322
  - as the postulate of law of relative quantities, 542–544
  - bin development pattern of, 609–611
  - lack of digital development of, 341–344
  - Leading Digits Inflection Point of, 347–348
  - summation along digits of, 366–370
  - quantitative reversal with, 515–517
  - universal law of relative quantities and, 626
- Last-Two Digits, 25, 26
- Leading Digit Inflection Point (LDIP), 345–348
- Leading Digit Parable, 200–206, 310–312, 313–316, 529
- Leemis, Lawrence, 273
- Leuenerger, Christoph, 482
- Logarithmic Distribution, 21, 27
- logarithmic-ness, 582–589, 617
- Lognormal Distribution, 272, 275–278, 325, 329, 330, 335–338, 507, 508, 615–618
  - development pattern of, 339, 340
  - in nature, 375, 376
  - Leading Digits Inflection Point of, 345–347
  - summations along digits of, 363, 364
- Ma, Bo-Qiang, 377
- Madoff, Bernard, 91
- Man Investments, 100
- mantissa, 253–260, 266, 441
- market capitalization, 96, 97
- Maxwell, James Clerk, 630
- measured categories, 497–500
- Mendeleev, Dmitri, 384
- Microsoft, 98, 99
- Miller, Steven, 226, 469, 470
- molecular mass of compounds, 381–385
- multiplication processes, 58–66
- multiplication table, 9–13, 316
- Multiplicative Central Limit Theorem (MCLT), 66, 277, 325, 333–334
  - in physical data, 389–394
  - in rock breaking, 400, 401
  - number system invariance of, 527
- NASA, 278
- NASDAQ, 89
- Net Asset Value (NAV), 90, 91, 100, 101
- New York City population, 42, 145, 332
- New York Stock Exchange, 9, 10, 89
- Newcomb, Simon, 15, 18, 19, 31, 630, 634
- Newton, Isaac, 394, 524, 525, 629
- Newton's Method, 241
- Nigrini, Mark, 80
- Normal Distribution, 290, 291, 304, 336
  - parametrical effects of, 474–477
  - products of, 394
- number system invariance principle, 519–521
- number systems, 234, 235, 236, 511

- Order of Magnitude (OOM), 33, 184, 186,  
187, 274, 281, 289, 575–577,  
585–587
- Order of Magnitude of Variability (OMV),  
34, 187, 274, 575–577,  
585–587
- payroll accounting data, 91, 95
- Periodic Table, 380, 381
- permutations (of balls & boxes), 408, 409
- physical constants set, 387, 388
- Pinkham, Roger, 230, 231, 499, 614
- planet and star formation model,  
415–418
- Ponzi schemes, 91
- population data, 41–47, 332–334,  
405, 406
- prime numbers, 241
- Proposition I, 261
- Proposition II, 262
- Proposition III, 263
- Proposition IV, 264
- Proposition V, 318
- Proposition VI, 318
- Quantitative Order of Magnitude (QTM),  
585–587
- quantitative reversal, 513–518
- quantum mechanics, 377–379, 413
- Raimi, Ralph, 230, 438
- Random Linear Combinations, 49–52,  
171, 177–191, 192–194, 380–386
- Random Walk, 65
- Rayleigh Distribution, 474–477
- rebellious exponential series (rates),  
427–438, 578, 619, 620
- Related Log Conjecture, 266–270,  
271–274, 282–292, 337, 617
- Hill's model and, 293, 294  
in rock breaking, 401  
scale invariance and, 295, 296  
superseding MCLT, 394
- rock breaking, 398–401
- Roman Numerals, 235, 623, 633–635
- Ross, Kenneth, 334, 393
- S — start (bin schemes), 538
- Sagan, Carl, 401
- Sambridge, Malcolm, 35–37
- Saville slope/algorithm, 134–137, 157,  
158, 484–496
- Scale Invariance Principle, 72, 73,  
230–232, 295, 296, 497–500
- Scheme A, 537, 538
- Scheme B, 537, 538
- Second-Generation Distributions (SGD),  
198, 199
- Second-Order Digits, 23, 297–300, 384,  
442–447, 602–607
- self-powered sequence, 241, 320
- Shao, Lijing, 377
- significand, 254–258
- Simple Averaging Schemes, 207–211
- small data sets, 82, 98, 160, 421, 422
- spikes, 83, 84
- Spinoza, Baruch, 41
- star distance data, 278
- Stigler's Law, 210, 233, 453
- stock market, 9–11
- Sum Percent Deviations (SPD), 155, 156
- Sum Percent Second Order Deviations  
(SPD2), 161
- Sum Squares Deviation (SSD), 128–133
- sum-invariant characterization of  
Benford's Law, 363–371, 587, 620, 621
- Summation Test, 141–148, 169, 170,  
363–371

- super exponential growth series, 439–441
- super random number, 222, 228
  
- Tau-Ceti-f (planet), 519–521, 524, 527, 532–534, 539, 540, 545, 591, 598, 599, 620, 628–634
- The Z-test, 120–122
- thermodynamics, 377–379, 402–413
- Third-Order Digits, 24, 442, 443
- total return, 90, 100, 101
- troughs, 83, 84
  
- U.S. County Area data, 172–176, 300, 360–362, 371, 576–580, 588, 589
- U.S. population centers, 42–45, 124, 130, 153, 161–163, 303, 327–329, 349–353, 405, 406, 484, 485, 520, 607, 614, 615
- unequal bin sizes, 599–601
- Uniform Distribution, 181, 182, 291, 292, 304, 393
- universal law of relative quantities, 626, 627, 632
  
- Value Repetition Test, 138–140, 170
- Varian, Hal, 79
  
- w — width (bin schemes), 536
- Wald Distribution, 479
- Weaver, Warren, 32
- Weibull Distribution, 479
  
- Zikuma's Law, 521, 537, 545, 591