

Cambridge University Press & Assessment

978-1-009-12307-5 — Applying Benford's Law for Assessing the Validity of Social Science Data

Michael A. Long , Paul B. Stretesky , Kenneth J. Berry ,

Janis E. Johnston , Michael J. Lynch

Frontmatter

[More Information](#)

APPLYING BENFORD'S LAW FOR ASSESSING THE VALIDITY OF SOCIAL SCIENCE DATA

Benford's law is a probability distribution for the likelihood of the leading digit in a set of numbers. This book seeks to improve and systematize the use of Benford's law in the social sciences to assess the validity of self-reported data. The authors first introduce a new measure of conformity to the Benford distribution that is created using permutation statistical methods and employs the concept of statistical agreement. In a switch from a typical Benford application, this book moves away from using Benford's law to test whether the data conform to the Benford distribution to using it to draw conclusions about the validity of the data. The concept of "Benford validity" is developed, which indicates whether a dataset is valid based on comparisons with the Benford distribution, and in relation to this, a diagnostic procedure is devised that assesses the impact on data analysis of not having Benford validity.

MICHAEL A. LONG is Professor of Sociology and Director of the Center for Insecurity and Inequality Research at Oklahoma State University, United States. He is the author or coauthor of six books and more than 90 journal articles and book chapters, primarily in the areas of environmental sociology, green criminology, sustainability, food insecurity, public health, and quantitative methodology. He has received funding for his research from the National Science Foundation, the US Department of Agriculture, the British Academy, and other institutions.

PAUL B. STRETESKY is Professor of Criminology at University of Lincoln, United Kingdom. He is the author of eight books and more than 100 journal articles and book chapters, primarily in the areas of environmental justice, environmental sociology, green criminology and food insecurity. His research has funding from the Natural Environment Research Council (UK).

KENNETH J. BERRY is Emeritus Professor of Sociology at Colorado State University, United States. He is the author of seven books and 185 referred journal articles. He is a leading scholar in the statistics

Cambridge University Press & Assessment

978-1-009-12307-5 — Applying Benford's Law for Assessing the Validity of Social Science Data

Michael A. Long , Paul B. Stretesky , Kenneth J. Berry ,

Janis E. Johnston , Michael J. Lynch

Frontmatter

[More Information](#)

subfield known as permutation methods. In 2002 he was named John N. Stern Distinguished Professor in the College of Liberal Arts at Colorado State University.

JANIS E. JOHNSTON works in the US Department of Agriculture (USDA) as a policy analyst. She is coeditor of a book on social equality, coauthor of five statistics books that focus on the subfield of statistics known as permutation methods and author of over 40 journal articles with a focus on quantitative methods.

MICHAEL J. LYNCH is Professor of Criminology at the University of South Florida, United States. He is the founder of the subfield of criminology known as green criminology. He is the author or editor of 20 books, some of which have multiple editions, more than 120 journal articles, and 55 book chapters.

Cambridge University Press & Assessment

978-1-009-12307-5 — Applying Benford's Law for Assessing the Validity of Social Science Data

Michael A. Long , Paul B. Stretesky , Kenneth J. Berry ,

Janis E. Johnston , Michael J. Lynch

Frontmatter

[More Information](#)

APPLYING BENFORD'S LAW FOR ASSESSING THE VALIDITY OF SOCIAL SCIENCE DATA

MICHAEL A. LONG

Oklahoma State University

PAUL B. STRETESKY

University of Lincoln

KENNETH J. BERRY

Colorado State University

JANIS E. JOHNSTON

US Government

MICHAEL J. LYNCH

University of South Florida



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press & Assessment

978-1-009-12307-5 — Applying Benford's Law for Assessing the Validity of Social Science Data

Michael A. Long, Paul B. Stretesky, Kenneth J. Berry,

Janis E. Johnston, Michael J. Lynch

Frontmatter

[More Information](#)



CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314-321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781009123075

DOI: 10.1017/9781009127950

© Michael A. Long, Paul B. Stretesky, Kenneth J. Berry, Janis E. Johnston, and
Michael J. Lynch 2024

This publication is in copyright. Subject to statutory exception and to the provisions
of relevant collective licensing agreements, no reproduction of any part may take
place without the written permission of Cambridge University Press & Assessment.

First published 2024

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

Library of Congress Cataloging-in-Publication Data

NAMES: Long, Michael A., author.

TITLE: Applying Benford's Law for assessing the validity of social science data / Michael A. Long,
Oklahoma State University, Paul B. Stretesky, Northumbria University, Kenneth J. Berry, Colorado
State University, Janis E. Johnston, USDA, Michael J. Lynch, University of South Florida.

DESCRIPTION: Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2024. |
Includes bibliographical references and index.

IDENTIFIERS: LCCN 2023016252 (print) | LCCN 2023016253 (ebook) | ISBN 9781009123075
(hardback) | ISBN 9781009124249 (paperback) | ISBN 9781009127950 (epub)

SUBJECTS: LCSH: Social sciences—Statistical methods. |

Distribution (Probability theory) | Quantitative research—Evaluation.

CLASSIFICATION: LCC HA29 .L83745 2024 (print) | LCC HA29 (ebook) |

DDC 300.72/7—dc23/eng/20230414

LC record available at <https://lcn.loc.gov/2023016252>

LC ebook record available at <https://lcn.loc.gov/2023016253>

ISBN 978-1-009-12307-5 Hardback

ISBN 978-1-009-12424-9 Paperback

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will
remain, accurate or appropriate.

Contents

<i>List of Figures</i>	<i>page</i> vii
<i>List of Tables</i>	ix
<i>Preface</i>	xiii
<i>Acknowledgments</i>	xv
1 Introduction	I
Benford's Law	6
Benford Agreement Analysis and Benford Validity	7
Plan for the Book	8
2 Validity and Self-Reported Data	II
Validity in Social Science	II
Validity of Self-Reported Data	15
Benford's Law and the Validity of Self-Reported Data	21
Conclusion	22
3 Benford's Law and Assessing Conformity	24
A Brief History	24
The Benford Distribution	28
Measures of Probability	36
Permutation Statistical Methods	41
Measures of Effect Size	46
A Chance-Corrected Measure of Effect Size	55
Ordinary Euclidean Scaling	58
Probability Values	63
Assessing Conformity	66
Conclusion	72
4 Data Characteristics and the Workflow of Benford Agreement Analysis	74
Data Characteristics and Assessing Accuracy	74
The Benford Distribution and Measures of Agreement	78
Program Benford	82

vi	<i>Contents</i>	
	Assessing Accuracy and Measuring Agreement: Recap	86
	The Workflow of Benford Agreement Analysis	86
	Conclusion	93
5	Benford Agreement Analysis of the Sea Around Us Project's Fish-Landings Data	94
	The Global Fishing Industry	94
	Sea Around Us and the Dataset	98
	Benford Agreement Procedure	101
	Benford Agreement Analysis of Sea Around Us Data	102
	Summary of the Sea Around Us Benford Agreement Analysis	128
	Conclusion	130
6	Benford Agreement Analysis of US and Global COVID-19 New Cases Data	131
	COVID-19, Background	132
	Benford Agreement Procedure	133
	The United States and COVID-19	133
	Benford Agreement Analysis of US COVID-19 New Cases Data	142
	Global COVID-19 Cases	149
	Conclusion	168
7	Assessing the Impacts of Problematic Benford Validity	169
	Assessing Data with Problematic Benford Validity	169
	Conclusion	186
8	Conclusion	188
	The Workflow of Benford Agreement Analysis: A Summary	188
	When to Use a Benford Agreement Analysis	194
	Concluding Thoughts	195
	<i>References</i>	196
	<i>Index</i>	204

Figures

3.1	Newcomb’s distributions of first and second significant digits.	<i>page 27</i>
3.2	Graphic of the observed and the expected probabilities for land areas of 196 countries.	34
3.3	Observed and expected frequencies for populations of 3,142 US counties.	36
3.4	Graphic of the observed and the expected probabilities for $2^n = 1,000$.	38
3.5	Graphic of Kuiper’s maximum D^+ and D^- distances.	54
4.1	Histogram of the observed data.	85
4.2	State prison population, $n = 1,479$.	89
5.1	Full Sea Around Us dataset, $n = 467,451$.	105
5.2	Full reported fish landings, $n = 256,814$.	106
5.3	Full unreported fish landings, $n = 210,637$.	106
5.4	1950 (reported), $n = 3,009$.	108
5.5	2010 (reported), $n = 4,206$.	108
5.6	1950 (unreported), $n = 2,596$.	108
5.7	2010 (unreported), $n = 3,344$.	109
5.8	West Africa (reported), $n = 26,053$.	115
5.9	West Africa (unreported), $n = 26,522$.	116
5.10	Graphic showing Simpson’s paradox.	118
5.11	Equatorial Guinea (reported), $n = 1,023$.	119
5.12	Gambia (reported), $n = 839$.	120
5.13	Gabon (unreported), $n = 606$.	121
5.14	Sierra Leone (unreported), $n = 1,104$.	122
5.15	Guinea (unreported), $n = 1,012$.	127

viii	<i>List of Figures</i>	
6.1	Daily US new COVID-19 cases, January 23, 2020–January 12, 2022.	136
6.2	Colombia, $n = 675$.	164
6.3	Venezuela, $n = 614$.	165

Tables

1.1	Benford's law: probability values for first, second, third, and fourth significant digits for $d = 0, \dots, 9$.	<i>page 6</i>
3.1	Newcomb's probability values for first and second significant digits for $d = 0, \dots, 9$.	26
3.2	Benford's counts of the average number of times the natural numbers 1, ..., 9 occur as first digits for 20 datasets.	29
3.3	Listing of digits from 1 to 300.	30
3.4	List of 20 countries by land area.	33
3.5	Frequency distribution of leading digits of 196 country land areas with expected frequencies under a Benford distribution.	34
3.6	List of 20 US counties by population size.	35
3.7	Frequency distribution of leading digits of 3,142 county populations with expected frequencies under a Benford distribution.	36
3.8	Listing of the first 20 occurrences of 2^n for $n = 1,000$, and leading digits.	37
3.9	Nine observed and theoretical expected probabilities of 2^n for $n = 1,000$.	37
3.10	Observed and expected probabilities for digits $d = 1, \dots, 6$.	40
3.11	Arrangements of $n = 10$ balls in $k = 5$ bins.	43
3.12	Comparisons of p and M for $1 \leq n = k \leq 20$.	44
3.13	Example dataset for comparisons of Pearson's chi-squared, Wilks' likelihood-ratio, and Fisher's exact goodness-of-fit tests.	44
3.14	Observed and expected probabilities for $d = 1, \dots, 9$.	50
3.15	Example data for the calculation of \mathfrak{N} with $k = 5$ categories.	60
3.16	Calculation of summations for μ_δ with $v = 2$ and $v = 1$.	61
3.17	Mendel's second-generation hybridization frequency data for $n = 556$ common garden peas.	65

x	<i>List of Tables</i>	
3.18	Observed and expected probabilities for $d = 1, \dots, 9$.	66
4.1	Chance-corrected measures of agreement for sample sizes 100 to 4,000 and for digits 3 to 6 in expected values.	75
4.2	MAD (δ) and \mathfrak{R} values for 2^n and the Fibonacci series for $n = 1, \dots, 1,000$.	76
4.3	Example data to illustrate the differences between measures of agreement and correlation.	81
4.4	First 40 cases of dataset Data1.txt.	82
4.5	State prison population, full dataset, 1972–2000, $n = 1,479$.	87
4.6	State prison population, full dataset, with observed minus Benford probabilities.	89
5.1	Example of Sea Around Us data.	99
5.2	Benford analysis for Sea Around Us fish-landings data, full dataset, and five different 10% random samples.	104
5.3	Benford analysis for Sea Around Us, full reported and unreported fish-landings data.	105
5.4	Benford analysis for Sea Around Us fish-landings data, by decade.	107
5.5	Benford analyses results for West African countries.	111
5.6	Data characteristics of West African countries.	113
5.7	Correlations between sample sizes, \mathfrak{R} , and orders of magnitude for Equatorial Guinea's reported and unreported fish-landings data.	114
5.8	West African reported fish landings, 1950–2016, $n = 26,053$.	115
5.9	West African unreported fish landings, 1950–2016, $n = 26,522$.	116
5.10	Equatorial Guinea's reported fish landings, 1950–2016, $n = 1,023$.	119
5.11	Gambia's reported fish landings, 1950–2016, $n = 839$.	120
5.12	Gabon's unreported fish landings, 1950–2016, $n = 606$.	121
5.13	Sierra Leone's unreported fish landings, 1950–2016, $n = 1,104$.	122
5.14	Observed probabilities and Benford agreement measures for random samples (RS) of reported and unreported fish landings of different sample sizes.	124
5.15	Guinea unreported fish landings, 1950–2016, $n = 1,012$.	127
6.1	Centers for Disease Control–confirmed COVID-19 cases, by US state and territory, January 22, 2020–January 12, 2022.	137

<i>List of Tables</i>	xi
6.2 State and territory rankings, top 15 by category.	140
6.3 US state and territory list (CDC reporting jurisdictions).	141
6.4 US COVID-19, cases and deaths by state, over time, US CDC dataset.	143
6.5 Full US CDC COVID-19, new cases data, $n = 32,819$.	144
6.6 CDC US COVID-19, new cases data, states/ territories, grouped alphabetically.	145
6.7 CDC COVID-19, new cases data by region of country.	146
6.8 CDC COVID-19, new cases data by day of the week.	147
6.9 US CDC COVID-19, new cases data by political party of governor.	149
6.10 Reported COVID-19 cases, by country, December 22, 2021, top 20 countries.	151
6.11 Reported COVID-19 cases, by country, December 22, 2021, lesser reporting countries.	153
6.12 Global COVID-19 cases and deaths by state, over time, OWID dataset.	155
6.13 Countries and characteristics.	156
6.14 Full global COVID-19, new cases data, $n = 105,703$.	158
6.15 Global COVID-19, new cases data, states/ territories grouped alphabetically.	159
6.16 Global COVID-19, new cases data by region.	161
6.17 Global COVID-19, new cases data by region of South America.	162
6.18 Global COVID-19, new cases data for some South American countries.	162
6.19 Global COVID-19, new cases data for Colombia, $n = 675$.	163
6.20 Global COVID-19, new cases data for Venezuela, $n = 614$.	164
6.21 Global COVID-19, new cases data by GFS.	166
7.1 Full aggregated reported fish landings, $n = 9,359$.	170
7.2 Countries in the analyses.	172
7.3 Fully aggregated reported fish landings (sample size ≥ 500), $n = 7,688$.	173
7.4 Example of cross-national dataset.	174
7.5 Descriptive statistics for variables in the Benford validity regression models for countries with \mathcal{N} values based on $n \geq 500$.	175
7.6 Random effects regression coefficients and robust standard errors for prediction of reported fish landings, landed values in millions of USD, 1960–2016.	178

xii	<i>List of Tables</i>	
7.7	Random effects regression coefficients and robust standard errors for the prediction of reported fish landings, landed values in millions of USD, 1960–2016, controlling for the level of Benford agreement (\mathcal{R}).	180
7.8	Random effects regression coefficients and robust standard errors for the prediction of reported fish landings, landed values in millions of USD, 1960–2016, controlling for the level of Benford agreement (\mathcal{R}) measured as a dichotomous variable.	182
7.9	Random effects regression coefficients and robust standard errors for the prediction of reported fish landings, landed values in millions of USD, 1960–2016 (only countries with acceptable agreement).	184
7.10	Binary logit regression coefficients and robust standard errors for the prediction of unacceptable agreement (\mathcal{R}).	185

Preface

Benford's law is a probability distribution for the leading digits in a set of natural numbers in which the nine leading digits are not equally likely (there are nine because zero is impermissible as a leading digit). Rather the leading digit probabilities range in descending order from 0.3010 for 1s to just 0.0458 for 9s. The Benford probability distribution has been occasionally used by mathematicians and other scientists to check the accuracy of the data with which they are working. In short, comparing observed data with the expected Benford distribution probabilities provides some information on the accuracy of the data. The closer the probability distribution of the observed data to the Benford probability distribution, the more accurate the data.

It may seem strange that a group of five sociologists, criminologists, and statisticians originally trained as sociologists decided to write a book about Benford's law. This is especially true, given that there are several books on Benford's law and its application. However, none of these existing works has come from the perspective of social scientists, or has a focus on the kinds of data that are frequently used by social scientists. Despite growing interest in the application of Benford's law, there is, to our knowledge, no book that provides a workflow for Benford analyses in the social sciences and software to conduct such analyses. The purpose of the current book is to fill in this gap. Our book also introduces a new statistical measure to assess conformity to the Benford distribution based on the statistical concept of "agreement" and uses permutation statistical methods rather than the more common frequentist approaches. The new measure – symbolized by the Fraktur letter \mathfrak{R} , the R computer program to calculate it (and related information), and the workflow we propose – can be used by social scientists to assess their data for agreement with the Benford distribution. Thus researchers will be provided with a quantitative measure of the validity of their data.

The idea for this book came from three of its authors (Mike Long, Paul Stretesky, and Mike Lynch) who, in their collaborative research, have carried out empirical work in environmental sociology, green criminology, and related areas for the past 20 years and have sometimes questioned the validity of the often self-reported data that are used in empirical studies of pollution and in related measures of environmental degradation. A search of the literature returned a handful of uses of Benford's law in assessing the accuracy of social science data. We then came across Mark Nigrini's book *Benford's Law: Applications for Forensic Accounting, Auditing and Fraud Detection* (2012) in the field of accounting. Nigrini's book is a well-written, comprehensive introduction to the use of Benford's law in applied research, with numerous examples from accounting. However, the main measure of conformity used by Nigrini, the mean absolute deviation (MAD), while better than the other measures used in the Benford literature, left some room for improvement.

After reviewing the literature on the use of Benford's law to assess the accuracy of social science data, Long realized that the measurement of conformity to the Benford distribution could be performed more efficiently using permutation statistical methods, an area of statistics he had worked on briefly in his early career. Long then recruited his previous collaborators – Kenneth Berry and Janis Johnston, two leading scholars of permutation statistical methods – to develop a permutation-based measure of statistical agreement for testing the observed data against the Benford distribution. The resulting team of five scholars decided it was time for a more systematic use of Benford's law to assess the validity of data used by social scientists. And thus the idea for this book was born.

Our primary motivation for writing this book is to give social scientists a method and a resource for analyzing the validity of self-reported (often secondary) social science data that meet certain conditions. The use of self-reported data in the social sciences is very common, and we believe that more rigor is needed in assessing their validity. A great deal of attention is given to diagnostics in statistical analysis; however, we believe that Benford's law can become a common diagnostic tool for data validity in the social sciences.

Acknowledgments

We would like to acknowledge a handful of individuals without whom this book and the ideas in it would not have come to be. First, we would like to give a huge thank you to David Repetto, Executive Publisher at Cambridge University Press, and to the Press in general, for giving us the opportunity to write this book. None of the authors has worked extensively with Benford's law in the past, but David and the Press showed faith in us that we could get the job done. Second, we would like to highlight Paul W. Mielke Jr.'s path-breaking work in permutation statistical methods. It is not only that his expertise and passion for these methods inspired the work of numerous authors of the present book, but without his earlier contributions in the area the main measure of validity of self-reported data we propose in this book would not have been possible. Third, we would like to acknowledge the work of W. S. Robinson, the sociologist whose name is most famously associated with the "ecological fallacy." In the 1950s Robinson wrote two articles in the journal *American Sociological Review* that detail the concept of agreement in statistics and develop the first measure of agreement, as an alternative to measures based on correlation. Our measure of validity employs agreement rather than correlation.

We would also like to thank the hundreds of colleagues and of graduate and undergraduate students with whom all of us have had conversations regarding the validity of self-reported data. It is these kinds of conversations and debates that help us become better social scientists and highlight the need to be rigorous in all the aspects of our research design, data collection, and data analysis. Finally and most importantly, we would like to thank our families, whose support and understanding is vital to completing a project like this.