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 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.

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

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

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://lccn.loc.gov/2023016252
LC ebook record available at https://lccn.loc.gov/2023016253

ISBN 978-1-009-12307-5 Hardback
ISBN 978-1-009-12424-9 Paperback

Contents

Figures

Tables

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.