Preface

One of the greatest beauties in mathematics is how the same equations can describe phenomena in widely different fields. Benford's Law of digit bias is an outstanding example of this. Briefly, it asserts that for many natural data sets we are more likely to see numbers with small leading digits than large ones. More precisely, our system is Benford if the probability of a first digit of d is $\log_{10} \frac{d+1}{d}$; we often consider the related but stronger statement that the probability of a significand being at most s is $\log_{10} s$, or the natural generalizations to other number bases. Base 10, the probabilities range from having a leading digit of 1 almost 30% of the time, to only about a 4.6% chance of starting with a 9.

Benford's Law arises in a variety of disciplines, including accounting, computer science, dynamical systems, economics, engineering, medicine, number theory, probability, psychology and statistics, to name just a few, and provides a wonderful opportunity for a common meeting ground for people with diverse interests and backgrounds. My first encounter with it was in Serre's *A Course in Arithmetic* [Ser]. On page 76 he remarks that Bombieri showed him a proof that the analytic density of the set of primes with leading digit 1 is $\log_{10} 2$, which is the Benford probability; a short argument using Poisson Summation yields the proof. I next saw it in Knuth's *The Art of Computer Programming, Volume 2: Seminumerical Algorithms* ([Knu], page 255), where he discusses applications of Benford's Law to analyzing floating point operations, especially the fact that Benford behavior implies the relative error from rounding is typically higher than one would expect. Once aware (or perhaps I should say doubly aware) of its existence, I saw it more and more often.

Our purposes here are to show students and researchers useful techniques from a variety of subjects, highlight the connections between the different areas and encourage research and cross-departmental collaboration on these problems. To do this, we develop much of the general theory in the first few chapters (concentrating on the methods which are applicable to a variety of problems), and then conclude with numerous chapters on applications written by world-experts in that field. Though there are common themes and methods throughout the applications, these chapters are self-contained, needing only the introductory chapters and some standard material. **For those wishing to use this as a textbook, numerous exercises and supplemental material are collected in the final chapter, and additionally are posted online (where more problems can easily be added, and links to relevant material for that chapter are collected); see**

http://web.williams.edu/Mathematics/sjmiller/public_html/benford/.

One advantage of posting problems online is that this need not be a static list, and thus please feel free to email suggestions for additional exercises.

Below we briefly outline the major themes of the book.

- **Part I: General Theory I: Basis of Benford's Law:** We begin our study of Benford's Law with a brief introduction by Miller in Chapter 1. We concentrate on the history and some possible explanations, and briefly discuss a few of the many applications and central questions in the field.

  While for many readers this level of depth suffices, the subject can (and should!) be built on firm foundations. We do this in Chapter 2, where Berger and Hill rigorously derive many results through the use of appropriate σ-algebras. There are many approaches to proving a system satisfies Benford's Law. One of the most important is the Fundamental Equivalence (also called the uniform distribution characterization), which says a system $\{x_n\}$ satisfies Benford's Law base B if and only if its logarithm modulo 1 (i.e., $y_n = \log_B x_n \bmod 1$) is uniformly distributed. In other words, in the limit, the probability the logarithm modulo 1 lies in a subinterval $[a, b]$ of $[0, 1]$ is just $b - a$. The authors describe this and additional characterizations of Benford's Law (including the scale-invariance characterization and the base-invariance characterization), and prove many deterministic and random processes satisfy Benford's Law, as well as discussing flaws of other proposed explanations (such as the spread distribution approach).

  For the uniform distribution characterization to be useful, however, we need ways to show these logarithms are uniformly distributed. Often techniques from Fourier analysis are well suited for such an analysis. The Fundamental Equivalence reduces the Benfordness of $\{x_n\}$ to the distribution of the fractional parts of its logarithms $\{y_n\}$. Fourier analysis is built on the functions $e_m(t) := \exp(2\pi imt)$ (where $i = \sqrt{-1}$); note that the painful modulo condition in y_n vanishes when it is the argument of e_m, as $e_m(y_n) = e_m(y_n \bmod 1)$. Chapter 3 by Miller is devoted to developing Fourier analytic techniques to prove Benford behavior. We demonstrate the power of this machinery by applying it to a variety of problems, including products and chains of random variables, L-functions, special densities and the infamous $3x + 1$ problem. For example, using techniques from Fourier analysis (especially Poisson Summation), one can show that the standard exponential random variable is very close to satisfying Benford's Law. The exponential is a special case of the three-parameter Weibull distribution. A similar analysis shows that, so long as the shape exponent of the Weibull is not too large, it too is close to being Benford. There are numerous applications of these results. The closeness of the standard exponential to Benford implies that order

statistics are almost Benford as well. The Weibull distribution arises in many survival models, and thus the analysis here provides another explanation of the prevalence of Benford behavior in many diverse systems.

- **Part II: General Theory II: Distributions and Rates of Convergence:** Combinations of data sets or random variables are often closer to satisfying Benford's Law than the individual data sets or distributions. This suggests a natural problem: looking for distributions that are exactly or at least close to being Benford. One of the most important examples of a distribution that exhibits Benford behavior is that of a geometric random variable. Numerous phenomena obey a geometric growth law; in particular, the solution to almost any linear difference equations is a linear combination of geometric series. We then investigate other important distributions and see how close they are to Benford. Although Benford's Law applies to a wide variety of data sets, none of the popular parametric distributions, such as the exponential and normal distributions, conforms exactly. Chapter 4 highlights the failure of several well-known probability distributions, then delves into the geometry associated with probability distributions that obey Benford's Law exactly. The starting point of these constructions is the fact that if U is a uniform random variable on $[a, a + n]$ for some integer n, then $T = 10^U$ is Benford base 10.

  As the exponential and Weibull distributions are not exactly Benford, it is important to obtain estimates on the size of the deviations. There are many ways to obtain such bounds. In Chapter 3 these bounds were obtained from Poisson Summation and the Fourier transform; in Chapter 5 Dümbgen and Leuenberger derive bounds from the total variation of the density (and its derivatives). These results are applied to numerous distributions, such as exponential, normal and Weibull random variables.

  This part concludes with Chapter 6 by Schürger. Earlier in the book we showed geometric Brownian motions are Benford. While processes such as the stock market were initially modeled by Brownian motions, such models have several defects, and current work must incorporate jumps and heavy tails. This leads to the study of Lévy Processes. These processes are described in detail, and their convergence to Benford behavior is shown. The techniques required are similar to those for geometric Brownian motion. On the other hand, the class of Lévy processes is much more general than just geometric Brownian motion, with applications in stochastic processes and finance; in particular, these and related processes model financial data, which has long been known to closely follow Benford's Law.

The final parts of this book deal with just some of the many applications of Benford's Law. Due to space constraints it is impossible to discuss all of the places Benford's Law appears. We have therefore chosen to focus on just a few situations, going for depth over breadth. We encourage the reader to peruse the many resources, such as the searchable online bibliography at [BerH2] or the large compilation [Hu], for a tour through additional areas to explore.

- **Part III: Applications I: Accounting and Vote Fraud:** Though initially an amusing observation about the distribution of digits in various data sets, since then Benford's Law has found numerous applications in many diverse fields. We briefly survey some of these. Probably the most famous application is to detecting tax fraud, though of course it is fruitfully used elsewhere too. We start in Chapter 7 with some of the basics of accounting, where Cleary and Thibodeau describe how Benford's Law can be integrated into business statistics and accounting courses. In particular, in the American Statistical Association's 2005 report *Guidelines for Assessment and Instruction in Statistics Education*, the following four goals (among others) are listed for what students should know after a first statistics course: (1) that variability is natural, predictable and quantifiable; (2) that random sampling allows results of surveys and experiments to be extended to the population from which the sample was taken; (3) how to interpret statistical results in context; (4) how to critique news stories and journal articles that include statistical information, including identifying what's missing in the presentation and the flaws in the studies or methods used to generate the information. The rest of the chapter shows how incorporating Benford's Law realizes these objectives.

  Chapter 8 by Nigrini describes one of the most important applications of Benford's Law: detecting fraud. Many diverse systems approximately obey the law, and thus deviations often indicate fraud. The chapter begins by examining some data sets that follow the law (tax returns, the 2000 census, stream flow data and accounts payable data), and concludes by showing how Benford's Law successfully detected fraud in accounts payable amounts, payroll data and corporate numbers (such as Enron).

  We continue with another important example where Benford's Law has successfully detected fraud. Chapters 9 by Mebane and 10 by Roukema discuss how Benford's Law can detect vote fraud; the first chapter develops tests based on the second digit and explores its use in practice, while the second concentrates on a recent Iranian election whose official vote counts were claimed to be invalid. .

- **Part IV: Applications II: Economics:** While there is no dearth of interesting topics to explore, we have chosen to devote this part of the book to economics because of the huge impact of recent events. A spectacular example of this is given by European Union (EU) policy, and the situation in Greece. We begin in Chapter 11 by Rauch, Göttsche, Brähler and Engel with a description of EU practices and data from several countries. As the stakes are high, there is enormous pressure to misreport statistics to avoid being hit with EU deficit procedures. We continue in Chapter 12 by Tödter with additional analysis, especially of published economics research papers. A surprisingly large proportion of first digits of regression coefficients and standard errors violate Benford's Law, in contrast to second digits. Routine applications of Benford tests would increase the efficiency of replication exercises and raise the risk of scientific misconduct. Another issue discussed is fitting data to a Generalized Benford Law, a topic Lee, Cho and Judge address in Chapter

17; both of these chapters deal with the issues facing the public arising from researchers falsifying data. We conclude this part with an analysis of data from the U.S. financial sector. The main finding is that Benford's Law fits the data from before the housing crisis well, but not the data afterwards.

- **Part V: Applications III: Sciences:** In previous chapters we discussed which distributions fit (and which don't fit) Benford's Law, as well as tests to detect fraud. In this part we take a different approach, and explore the psychology behind the people generating numbers. Chapters 14 by Burns and Krygier and 15 by Chou, Kong, Teo and Zheng explore patterns and tendencies in number generation, and the resulting implications, followed by Hoyle's chapter on the prevalence of Benford's Law in the natural sciences, including a summary of its occurrences and a discussion of the consequences. We end in Chapter 17 by Lee, Cho and Judge with a nice mix of theory and application. The authors consider a generalization of Benford's Law, developing the theory and analyzing known cases of fraud. They study the related Stigler distribution, and describe how it may be found from information-theoretic methods. This leads to alternative digit distributions based on maximum entropy principles. The chapter ends by using these new distributions in an analysis of some medical data which was known to be falsified, where the falsified data is detected. An important application of the material of this part is in developing tests to detect whether researchers are submitting fraudulent data. Similar to the chapters from economics, as the costs to society from incorrectly adopting conclusions of faulty research can be high, these tests provide a valuable tool to check the veracity of claims.

- **Part VI: Applications IV: Images:** Our final part deals with whether or not images follow Benford's Law. Chiverton and Wells, in Chapter 18, explore the relationship between intensities in medical images and Benford behavior. They describe a simple classifier based on Bayes theory which uses the Benford Partial Volume (PV) distribution as a prior; the results show experimentally that the Benford PV distribution is a reasonable modeling tool for the classification of imaging data affected by the PV artifact. The fraud-based applications of Benford's Law have grown from financial data sets to others as well. The last chapter, Chapter 19 by Pérez-González, Quach, Abdallah, Heileman and Miller, explores whether or not Benford's Law can detect modifications in images. Specifically, while images in the pixel domain are not close to Benford, the result after applying the Discrete Cosine Transform is. These results can be used to look for hidden messages in pictures, as well as to test whether or not the image has been compressed.

We are extremely grateful to Princeton University Press, especially to our editor Vickie Kearn and to Betsy Blumenthal and Jill Harris, for all their help and aid, to our copyeditor Alison Durham who did a terrific job, especially in standardizing the exposition across chapters, to Meghan Kanabay for assistance with many of

the illustrations, and Amanda Weiss for help with the jacket design. Many people proofread the book, looking not just for grammatical issues but also making sure it was a coherent whole with widely accessible expositions; it is a pleasure to thank them, especially John Bihn and Jaclyn Porfilio.

The editor was partially supported by NSF Grants DMS0600848, DMS0970067 and DMS1265673; some of his students assisting with the project were supported by NSF Grants DMS0850577 and DMS1347804, the Clare Boothe Luce Program, and Williams College. Some of this book is based on a conference organized by Chaouki T. Abdallah, Gregory L. Heileman, Steven J. Miller and Fernando Pérez-González and assisted by Ted Hill: *Conference on the Theory and Applications of Benford's Law* (16–18 December 2007, Santa Fe, NM). This conference was supported in part by Brown University, IEEE, NSF Grant DMS-0753043, the New Mexico Consortium's Institute for Advanced Study, Universidade de Vigo and the University of New Mexico, and it is a pleasure to thank them and the participants.

Steven J. Miller
Williams College
Williamstown, MA
October 2013

sjm1@williams.edu, Steven.Miller.MC.96@aya.yale.edu