Benford’s Law and Detection of Anomalies in Data

Dr. Ted Hill
School of Mathematics, Georgia Tech
California Polytechnic State University

Outline

- Brief History of Benford’s Law (BL)
- Use of BL to Detect Anomalies in Data
  - Fraud
  - Other anomalies
- Seven Basic BL Probability Theorems
- Common Errors related to BL
- How to win € from your friends

Benford’s Law for First Digits

Prob (First digit of \( X \) is \( d \)) = \( \log_{10}(1+d^{-1}) \), \( d = 1,2,\ldots,9 \)

i.e.,
\[
\begin{align*}
P(D_1(X) = 1) &= \log_{10}(2) \approx .301 \\
P(D_1(X) = 2) &= \log_{10}(1.5) \approx .176 \\
\vdots \\
P(D_1(X) = 9) &= \log_{10}(1+0.111\ldots) \approx .046
\end{align*}
\]

(Here \( D_i \) is the first significant digit (base 10) of \( x > 0 \).

e.g., \( D_1(2019) = D_i(0.02019) = 2 \))
**First-digit Dataset (Benford 1938)**

<table>
<thead>
<tr>
<th>Group</th>
<th>Title</th>
<th>First Digit</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Rivers, Area</td>
<td>31.0</td>
<td>335</td>
</tr>
<tr>
<td>B</td>
<td>Populations</td>
<td>100.0</td>
<td>335</td>
</tr>
<tr>
<td>C</td>
<td>Common</td>
<td>31.5</td>
<td>335</td>
</tr>
<tr>
<td>D</td>
<td>Newspapers</td>
<td>100.0</td>
<td>335</td>
</tr>
<tr>
<td>E</td>
<td>Spec. Issues</td>
<td>100.0</td>
<td>335</td>
</tr>
<tr>
<td>F</td>
<td>Distributions</td>
<td>100.0</td>
<td>335</td>
</tr>
<tr>
<td>G</td>
<td>Mol. Wgt.</td>
<td>100.0</td>
<td>335</td>
</tr>
<tr>
<td>H</td>
<td>ABNCAge</td>
<td>100.0</td>
<td>335</td>
</tr>
<tr>
<td>I</td>
<td>ABNCWgt.</td>
<td>100.0</td>
<td>335</td>
</tr>
<tr>
<td>J</td>
<td>Others</td>
<td>100.0</td>
<td>335</td>
</tr>
</tbody>
</table>

**Empirical Evidence of BL Today**

- exact first-digit law ($\Delta = 0$)
- Benford's combined data ($\Delta = 0.89$)
- populations of U.S. counties ($\Delta = 1.41$)
- numbers on WWW ($\Delta = 1.56$)
- US tax returns ($\Delta = 0.48$)

**BL Fraud Detection**

(Key Idea by M. Nigrini 1990's)

- Tax (individual, corporate, governmental)
- Clinical and drug trials
- Survey data
- Environmental
- Voting
- Health Insurance
- Scientific papers
- Fingerprint forgeries

**Benford's Own Data?**

<table>
<thead>
<tr>
<th>Group</th>
<th>Title</th>
<th>First Digit</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Rivers, Area</td>
<td>31.0</td>
<td>335</td>
</tr>
<tr>
<td>B</td>
<td>Populations</td>
<td>100.0</td>
<td>335</td>
</tr>
<tr>
<td>C</td>
<td>Common</td>
<td>31.5</td>
<td>335</td>
</tr>
<tr>
<td>D</td>
<td>Newspapers</td>
<td>100.0</td>
<td>335</td>
</tr>
</tbody>
</table>

**Diaconis & Freedman 1979:**

18 rounds to 5.4% and 19 rounds to 5.7%
BL Other Anomaly Detection – Phase Transitions

- Earthquakes (depths, time intervals)
- Quantum processes (many-body problems)

(Sambridge et al. 2011)
(Sen and Sen 2011)

BL Other Anomaly Detection - Image Processing

- Spectroscopic analysis (e.g., MRI’s)
- Steganography (hidden images)
- Natural vs. artificial images
- Image alterations

BL Other Anomaly Detection

- Internet traffic (intrusions, intentional & not)
- Music analysis (natural vs. artificially created chords)
- Sport game manipulation (detect match-fixing)
- Macroeconomics (GDP, purchasing power parity)
- Cardiology (different types of arrhythmia)

Related Application – Model Testing

- Math Model (differential equations, Monte Carlo, etc.)

Benford-In, Benford-Out Test

- 2010 Census (1990, 2000, 2010 all followed BL)
- 2050 Prediction
Seven Basic BL Probability Theorems

Thm 1. BL is the unique scale-invariant probability distribution on significant digits.
Ex. If a financial dataset X is Benford in €, it is also B in $.
If X is not Benford in € it is also not Benford in $
Ex. If distances to galaxies in light years follow BL, they will also follow BL measured in inches, centimeters, miles, and every other unit.

Thm 2. BL is the unique continuous base-invariant probability distribution on significant digits.

Thm 3. BL is the unique sum-invariant probability distribution on significant digits (Nigrini, Allaart).

BL Probability Theorems (cont’d)

Thm 4. If X is a Benford random variable, then so are $X^2$, $1/X$, and $XY$, where Y is any positive random variable independent of X.
Ex. If a financial dataset X is Benford in € per stock, it is also Benford in stock per €.
Ex. If $X_1 \times X_2 \times X_3 \times X_4 \times \ldots \times X_n$ are independent positive random variables (e.g. interest rates), then if any $X_i$ is Benford, then the whole product is Benford and remains Benford forever.

BL Probability Theorems (cont’d)

Thm 5. If X is a random variable with a density, then $X$, $X^2$, $X^3$, $X^4$, … is Benford with probability 1. (Berger-H).

Thm 6. If $X_1, X_2, X_3, X_4, \ldots$ are i.i.d. random variables with a density, then $X_1, X_2, X_3, X_4, \ldots$ is Benford with probability 1. (Berger-H).

Mixing Data from Different Distributions

Thm 7. Combining random samples from unbiased random distributions yields a Benford distribution in the limit (with probability 1).

Ex. Average ..... 10.6 18.5 12.4 9.4 8.0 6.8 5.1 4.8 4.5 4.2
Probable Error | 0.6 | 0.4 | 0.3 | 0.2 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0

BL Probability Theorems (cont’d)
**Three Common Errors**

1. *Not all* exponential sequences $a, a^2, a^3, ...$ are Benford.  
   **Ex.** If $a = \sqrt{10}$,  
   then the first digits of $a, a^2, a^3, ...$ are 3,1,3,1,3,1,...

2. *No* sequence $a, 2a, 3a, 4a, ...$ (or sums of iid random variables) are Benford.

3. A BL distribution need *not* cover many orders of magnitude.
   **Ex.** If $U$ is a Uniform(0,1) random variable, then  
   $X = 10^U$ is exactly Benford, and $1 \leq X < 10$.

**A Widespread Error**

4. Regularity and large spread do *not* imply BL.

---

**Online Resources**

- **Free searchable Benford Online Bibliography:**  
  [http://www.benfordonline.net/](http://www.benfordonline.net/)

- **Open-access monograph:** *A basic theory of Benford's law*  
  (Berger-H, 2011, Probability Surveys 8, 1-126)  
  [http://www.i-journals.org/ps/viewissue.php?id=11#Articles](http://www.i-journals.org/ps/viewissue.php?id=11#Articles)

- **Mathworld:**  

---

**Thank you, European Commission!**

And especially the organizers:

- **Domenico Perrotta**, European Commission,  
  Joint Research Centre, Italy (Chair)

- **Andrea Cerioli**, Università di Parma, Italy

- **Lucio Barabesi**, Università di Siena, Italy
Players I and U each choose a positive integer. Let $X = \text{product of the two integers}$.  
I win if $X$ begins with 1, 2, 3  
U win if $X$ begins with 4, 5, 6, 7, 8, or 9

We play 20 times – winner gets €10 from loser each time.