Breaking the Law (of Leading Digits): Why We Fail at Committing Fraud

Allison Lewis
University of Portland

NUMS Conference
April 9th, 2011
Introduction to Benford’s Law
- Statement of Law
- History
- Benford Tests
- Issues in Benford Analysis

Applications of Benford’s Law:
- Hydrology Statistics
- Iranian Election Results of 2009
- Climategate Data

Theory of Benford’s Law:
- Weibull Distribution
For any positive number $x$, we can write $x$ in scientific notation as

$$x = M_B(x) \cdot B^{k(x)}.$$

- $M_B(x)$ is called the **mantissa** of $x$

- $k(x)$ is an integer value which represents the exponent
Benford’s Law of Leading Digits

For many real-life data sets, the probability of observing a first digit of \( d \) base \( B \) is \( \log_B(1 + \frac{1}{d}) \).

In other words, the leading digits of most data sets are logarithmically, rather than uniformly, distributed.

Benford’s Law (Generalized)

The probability of observing a mantissa of at most \( s \) is \( \log_B s \).
For a Benford base 10 data set, we expect the leading digits to (approximately) follow the proportions below:

<table>
<thead>
<tr>
<th>Leading Digit</th>
<th>Benford Base 10 Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.30103</td>
</tr>
<tr>
<td>2</td>
<td>0.17609</td>
</tr>
<tr>
<td>3</td>
<td>0.12494</td>
</tr>
<tr>
<td>4</td>
<td>0.09691</td>
</tr>
<tr>
<td>5</td>
<td>0.07918</td>
</tr>
<tr>
<td>6</td>
<td>0.06695</td>
</tr>
<tr>
<td>7</td>
<td>0.05799</td>
</tr>
<tr>
<td>8</td>
<td>0.05115</td>
</tr>
<tr>
<td>9</td>
<td>0.04576</td>
</tr>
</tbody>
</table>
We can extend these proportions as far out into the mantissa as we wish:

<table>
<thead>
<tr>
<th>Digit</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.11968</td>
<td>0.10178</td>
<td>0.10018</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.30103</td>
<td>0.11389</td>
<td>0.10138</td>
<td>0.10014</td>
</tr>
<tr>
<td>2</td>
<td>0.17609</td>
<td>0.10882</td>
<td>0.10097</td>
<td>0.10010</td>
</tr>
<tr>
<td>3</td>
<td>0.12494</td>
<td>0.10433</td>
<td>0.10057</td>
<td>0.10006</td>
</tr>
<tr>
<td>4</td>
<td>0.09691</td>
<td>0.10031</td>
<td>0.10018</td>
<td>0.10002</td>
</tr>
<tr>
<td>5</td>
<td>0.07918</td>
<td>0.09668</td>
<td>0.09979</td>
<td>0.09998</td>
</tr>
<tr>
<td>6</td>
<td>0.06695</td>
<td>0.09337</td>
<td>0.09940</td>
<td>0.09994</td>
</tr>
<tr>
<td>7</td>
<td>0.05799</td>
<td>0.09035</td>
<td>0.09902</td>
<td>0.09990</td>
</tr>
<tr>
<td>8</td>
<td>0.05115</td>
<td>0.08757</td>
<td>0.09864</td>
<td>0.09986</td>
</tr>
<tr>
<td>9</td>
<td>0.04576</td>
<td>0.08500</td>
<td>0.09827</td>
<td>0.09982</td>
</tr>
</tbody>
</table>
Why Do We Care About Benford’s Law?
Benford’s Law can be used to demonstrate consistency in natural data sets (measured by conformance to the expected leading digit probabilities). Conversely, inconsistent results obtained from applying Benford tests to a data set may suggest the possibility of rounding errors or discrepancies in data collection methods, or even the presence of fraud or other data integrity issues.
First and Last Digit Tests:

- **First Digit**
  - $P(d_1) = \log_B(1 + \frac{1}{d_1})$

- **First Two Digits**
  - $P(d_1d_2) = \log_B(1 + \frac{1}{10d_1+d_2})$

- **First Three Digits**
  - $P(d_1d_2d_3) = \log(1 + \frac{1}{100d_1+10d_2+d_3})$

- **Last Digit**
  - $P(\text{last digit } d) = \frac{1}{10}$
Benford Tests (continued)

Last Two-Digit Tests:

- **All Endings**
  - \( P(\text{any ending } d_1 d_2) = \frac{1}{100} \)

- **Non-Doubles vs. Doubles**
  - \( P(\text{non-double}) = \frac{9}{10}, P(\text{double}) = \frac{1}{10} \)

- **Non-Doubles vs. Doubles (Split)**
  - \( P(\text{non-double}) = \frac{9}{10}, P(\text{any double } d_1 d_1) = \frac{1}{100} \)

- **Doubles (Conditional)**
  - \( P(d_1 d_1 | \text{double}) = \frac{1}{10} \)
Issues Arising in a Benford Analysis

- **Chi-square sensitivity to large data sets**
  - Alternative: Mean absolute deviation

- **Potentially non-Benford behavior**
  - Size of data set
  - Span of data set
  - Number of significant digits
Intentions

Discrepancies from Benford’s Law need not necessarily indicate fraud. It is *not* our intent to accuse anyone of such behavior! Our goal is to see whether or not certain data sets follow Benford’s Law and comment on the results.
Streamflow Data Set

- **Data Description**
  - Source: US Geological Survey
  - Spans time period of 130 years
  - Methods of data collection consistent

- **Characteristics**
  - Size: 457,440 data entries
  - Span: 9 orders of magnitude
  - Significant Digits: 3 or more
In this study, we analyze the first three digits.

Recall, the probability of observing a mantissa that begins with $d_1d_2d_3$ is:

$$\log_{10} \left( 1 + \frac{1}{100d_1 + 10d_2 + d_3} \right)$$
Approach: Remove all data entries with fewer than four significant digits to avoid counting rounded values.

Result: Limits data set to 73,828 values (16.1% of original data set), spanning only one order of magnitude (results in a strange, non-Benford distribution).

For comparison, we also ran tests on the complete data set with no restrictions.
Restricted Versus Unrestricted

**Figure:** (Left: Restricted; Right: Unrestricted) Comparing the First Three Digits Tests
### Test Results:

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Digit</td>
<td>45.82</td>
<td>0.00086</td>
</tr>
<tr>
<td>First 2 Digits</td>
<td>178.74</td>
<td>0.00017</td>
</tr>
<tr>
<td>First 3 Digits (Restricted)</td>
<td>12054.70</td>
<td>0.00039</td>
</tr>
<tr>
<td>First 3 Digits (Unrestricted)</td>
<td>23345.30</td>
<td>0.00020</td>
</tr>
</tbody>
</table>

**Table:** Starting Digit Tests: Hydrology Data
## Comparing Benford Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Restricted</th>
<th>Unrestricted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of Data Set</td>
<td>73,828</td>
<td>446,055</td>
</tr>
<tr>
<td>Orders of Magnitude</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td># Significant Digits</td>
<td>≥ 4</td>
<td>≥ 3</td>
</tr>
<tr>
<td>Mean Absolute Deviation</td>
<td>0.00039</td>
<td>0.00020</td>
</tr>
</tbody>
</table>

### Hydrology Conclusions

- Increasing size and span of data results in a better fit to the Benford distribution.
2009 Iranian Election

- Controversial presidential election in 2009
  - Allegations of ballot-stuffing fraud

- Previous Benford Tests:
  - Walter Mebane (2009) - Second Digit Analysis

- Polling vs. Precinct level data
  - Polling: 45,692 observations for each candidate
  - Precinct: 320 observations for each candidate
Polling Level Statistics

<table>
<thead>
<tr>
<th>Test</th>
<th>Total</th>
<th>Ahmadinejad</th>
<th>Mousavi</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Digit</td>
<td>3112.31</td>
<td>4121.17</td>
<td>366.32</td>
</tr>
<tr>
<td>Last Digit</td>
<td>398.87</td>
<td>11.82</td>
<td>7.63</td>
</tr>
<tr>
<td>Endings</td>
<td>2652.48</td>
<td>94.74</td>
<td>560.24</td>
</tr>
<tr>
<td>Non/Doubles</td>
<td>369.58</td>
<td>0.33</td>
<td>0.16</td>
</tr>
<tr>
<td>Non/Doubles(S)</td>
<td>2405.19</td>
<td>13.63</td>
<td>58.80</td>
</tr>
<tr>
<td>Doubles(C)</td>
<td>1603.09</td>
<td>13.19</td>
<td>58.31</td>
</tr>
</tbody>
</table>

**Table:** Polling Level - 45,692 observations
## Precinct Level Statistics

<table>
<thead>
<tr>
<th>Test</th>
<th>Total</th>
<th>Ahmadinejad</th>
<th>Mousavi</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Digit</td>
<td>24.84</td>
<td>14.80</td>
<td>6.41</td>
</tr>
<tr>
<td>Last Digit</td>
<td>12.44</td>
<td>3.81</td>
<td>4.88</td>
</tr>
<tr>
<td>Endings</td>
<td>104.38</td>
<td>96.88</td>
<td>94.38</td>
</tr>
<tr>
<td>Non/Doubles</td>
<td>0.31</td>
<td>2.81</td>
<td>0.14</td>
</tr>
<tr>
<td>Non/Doubles(S)</td>
<td>13.59</td>
<td>8.09</td>
<td>10.89</td>
</tr>
<tr>
<td>Doubles(C)</td>
<td>12.14</td>
<td>4.12</td>
<td>10.12</td>
</tr>
</tbody>
</table>

**Table:** Precinct Level - 320 observations
Election Data Approach

- Introduce randomness
- Test polling level data in subsets of 300 data entries
- Analyze averages of chi-square values from data subsets
### Chi-Square Averages: Polling Level (Split)

<table>
<thead>
<tr>
<th>Test</th>
<th>Total</th>
<th>Ahmadinejad</th>
<th>Mousavi</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Digit</td>
<td>29.14</td>
<td>36.84</td>
<td>9.92</td>
</tr>
<tr>
<td>Last Digit</td>
<td>11.24</td>
<td>8.71</td>
<td>9.10</td>
</tr>
<tr>
<td>Endings</td>
<td>114.88</td>
<td>99.93</td>
<td>102.17</td>
</tr>
<tr>
<td>Non/Doubles</td>
<td>3.47</td>
<td>0.99</td>
<td>1.03</td>
</tr>
<tr>
<td>Non/Doubles(S)</td>
<td>27.74</td>
<td>10.23</td>
<td>10.53</td>
</tr>
<tr>
<td>Doubles(C)</td>
<td>18.82</td>
<td>9.13</td>
<td>9.33</td>
</tr>
</tbody>
</table>

**Table**: Chi-Square Means: Polling Level (Split)
Conclusions

- Other possible factors: higher voter turn-out, growth in support for Ahmadinejad, increased turn-out from a previously silent majority
- Voter turn-out increased by 75% from previous presidential election
- Two provinces reported turn-out greater than 100%
Climategate Scandal

- **Massive E-mail leak at CRU - November 2009**
  - Allegations of scientific misconduct in the climate science community

- **Researchers Phil D. Jones and Michael E. Mann**
  - “Proxy Temperature Reconstruction" data from “Global Surface Temperatures Over the Past Two Millenia"

- **Data set with 32,451 observations**
  - Contains 30 data subsets covering different regions
  - Data entries measured as deviations from baseline temperature.
Problem:
Amalgamation of all thirty data subsets gave spike of values ending in 77 and deficit of values ending in 00.

Figure: Double-digit ending combinations in climate data
Approach

Analyze subsets of data with strange last two digit distributions:

- "Western US Unsmoothed" Data Set (1781 entries)
- "Tasmania Unsmoothed" Data Set (1991 entries)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>00</th>
<th>11</th>
<th>22</th>
<th>33</th>
<th>44</th>
<th>55</th>
<th>66</th>
<th>77</th>
<th>88</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>West. US</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>Tasmania</td>
<td>57</td>
<td>80</td>
<td>64</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table:** Ending Double-Digit Occurrences in Select Data Series
“Tasmania" Analysis

- 46 ending combinations not observed at all

- Range: [-4.43, 3.59]

<table>
<thead>
<tr>
<th>00</th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>08</th>
<th>09</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>0</td>
<td>0</td>
<td>72</td>
<td>2</td>
<td>0</td>
<td>79</td>
<td>0</td>
<td>49</td>
<td>2</td>
<td>0</td>
<td>80</td>
</tr>
</tbody>
</table>

**Table:** First 12 Ending Digit Occurrences for Tasmania Unsmoothed


Table: “Tasmania Unsmoothed" Data: Last Two Digits Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>Mean Abs. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endings</td>
<td>3261.49</td>
<td>0.0113</td>
</tr>
<tr>
<td>Non/Doubles</td>
<td>19.36</td>
<td>0.0296</td>
</tr>
<tr>
<td>Non/Doubles(S)</td>
<td>538.58</td>
<td>0.0163</td>
</tr>
<tr>
<td>Doubles(C)</td>
<td>400.68</td>
<td>0.1200</td>
</tr>
</tbody>
</table>
Conclusion

- Discrepancies should smooth out in an amalgamation of all 32,451 data entries
- Other potential factors: rounding errors, inconsistency in data collection techniques, or simply non-Benford behavior
Open Problem
Which probability distributions conform to Benford’s Law?

Outline
- Discuss relevance of the Weibull distribution in real-life situations
- Determine conformity of a random variable with a Weibull distribution
- Measure deviations depending on changing parameter values
Weibull Distribution

**Weibull Density Function**

\[
f(x; \gamma, \alpha, \beta) = \frac{\gamma}{\alpha} \cdot \left(\frac{x-\beta}{\alpha}\right)^{(\gamma-1)} \cdot e^{-\left(\frac{x-\beta}{\alpha}\right)^\gamma}
\]

\[x \geq \beta; \quad \gamma, \alpha > 0\]

- **Weibull Facts:**
  - Special cases include Exponential ($\gamma = 1$) and Rayleigh ($\gamma = 2$)
  - Used in survival analysis ($X$ represents "time-to-failure")
  - Models real-life data in engineering, medicine, politics, pollution, and numerous other fields
If a data set satisfies Benford’s Law, then its logarithms are uniformly distributed.

Benford’s Law is equivalent to saying $F_B(z) = z$, implying that our random variable is Benford if $F'_B(z) = 1$. Therefore, a natural way to investigate deviations from the Benford distribution is to compare the deviation of $F'_B(z)$ from 1, which would represent a uniform distribution.
Theorem (Miller, Cuff, and Lewis - 2010)

Let $Z_{\alpha,0,\gamma}$ be a random variable whose density is a Weibull with parameters $\beta = 0$ and $\alpha, \gamma > 0$ arbitrary. For $z \in [0, 1]$, let

$$F_B(z) := \text{Prob}(\log_B Z_{\alpha,0,\gamma} \mod 1 \in [0, z)).$$

1. The density of $Z_{\alpha,0,\gamma}$, $F'_B(z)$, is given by

$$F'_B(z) = 1 + 2 \sum_{m=1}^{M-1} \text{Re} \left[ e^{-2\pi im \left( z - \frac{\log \alpha}{\log B} \right)} \cdot \Gamma \left( 1 + \frac{2\pi im}{\gamma \log B} \right) \right]$$

$$+ \mathcal{E} \left( \frac{2\sqrt{2}M}{\pi^3} (40 + \pi^2) \sqrt{\gamma \log B} \cdot e^{-\pi^2 M / \gamma \log B} \right).$$
Theorem (continued)

2 For $m \geq \frac{\gamma \log B \log 2}{4\pi^2} \geq M$, the error from keeping the first $M$ terms is

$$|\mathcal{E}| \leq \frac{1}{\pi^3} 2\sqrt{2} M (40 + \pi^2) \sqrt{\gamma \log B} \cdot e^{-\pi^2 M/\gamma \log B}.$$

3 In order to have an error of at most $\epsilon$ in evaluating $F_B'(z)$, it suffices to take the first $M$ terms, where

$$M = \frac{k + \ln k + \frac{1}{2}}{a},$$

with $k \geq 6$ and

$$k = -\ln \left( \frac{a\epsilon}{C} \right), \quad a = \frac{\pi^2}{\gamma \log B}, \quad C = \frac{2\sqrt{2}(40 + \pi^2) \sqrt{\gamma \log B}}{\pi^3}.$$
**Figure:** $\gamma \in [0, 15]$. As $\gamma$ increases, the Weibull distribution is no longer a good fit compared to the uniform. Note that $\alpha$ has less of an effect on the overall conformance.
Figure: $\gamma \in [0, 2]$. As $\gamma$ increases, the Weibull distribution is no longer a good fit compared to the uniform. Note that $\alpha$ has less of an effect on the overall conformance.
Statement
Recall: If a data set satisfies Benford’s Law, then its logarithms are uniformly distributed.

We take the derivative of the CDF of the logarithms modulo 1 and compare it to the uniform distribution to calculate the deviation from Benford’s Law.
Techniques Used

Poisson Summation and Fourier Transform

As long as a function $H(k)$ is rapidly decaying, we may apply Poisson Summation, thus

$$\sum_{k=-\infty}^{\infty} H(k) = \sum_{k=-\infty}^{\infty} \hat{H}(k)$$

where $\hat{H}$ is the Fourier Transform of $H$:

$$H : \hat{H}(u) = \int_{-\infty}^{\infty} H(t)e^{-2\pi itu} dt.$$
Proof of Theorem (Part 1)

Let \( \zeta \) be a Weibull distribution with \( \beta = 0 \) and \([a, b] \subset [0, 1]\).

\[
F_B(b) = \text{Prob}(\log_B \zeta \mod 1 \in [0, b])
\]

\[
= \sum_{k=-\infty}^{\infty} \text{Prob}(\log_B \zeta \in [0 + k, b + k])
\]

\[
= \sum_{k=-\infty}^{\infty} \left( e^{-\left(\frac{B^k}{\alpha}\right)^\gamma} - e^{-\left(\frac{B^{b+k}}{\alpha}\right)^\gamma} \right)
\]
Proof of Theorem (Part 1 - continued)

\[ F'_B(b) = \sum_{k=-\infty}^{\infty} \frac{1}{\alpha} \cdot \left[ e^{-\left( \frac{B^{b+k}}{\alpha} \right)^\gamma} B^{b+k} \left( \frac{B^{b+k}}{\alpha} \right)^{-\gamma-1} \gamma \log B \right] \]

\[ = \sum_{k=-\infty}^{\infty} \frac{1}{\alpha} \cdot \left[ e^{-\left( \frac{ZB^k}{\alpha} \right)^\gamma} ZB^k \left( \frac{ZB^k}{\alpha} \right)^{-\gamma-1} \gamma \log B \right] \]

(where for \( b \in [0, 1] \), let \( Z = B^b \))

\[ = \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\alpha} \cdot e^{-\left( \frac{ZB^k}{\alpha} \right)^\gamma} ZB^k \left( \frac{ZB^k}{\alpha} \right)^{-\gamma-1} \gamma \log B \cdot e^{-2\pi itk} dt \]
Proof of Theorem (Part 1 - continued)

We use another change of variables:

\[ w = \left( \frac{\zeta B^t}{\alpha} \right)^\gamma \quad \text{or} \quad t = \log_B \left( \frac{\alpha w^{1/\gamma}}{\zeta} \right), \tag{1} \]

We can now use the gamma function:

\[
F'_B(z) = \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-w} \cdot \exp \left( -2\pi ik \cdot \log_B \left( \frac{\alpha w^{1/\gamma}}{\zeta} \right) \right) \, dw
\]

\[
= \sum_{k=-\infty}^{\infty} \left( \frac{\alpha}{\zeta} \right)^{-2\pi ik / \log B} \int_{-\infty}^{\infty} e^{-w} \cdot w^{-2\pi ik / \gamma \log B} \, dw
\]

\[
= \sum_{k=-\infty}^{\infty} \left( \frac{\alpha}{\zeta} \right)^{-2\pi ik / \log B} \Gamma \left( 1 - \frac{2\pi ik}{\gamma \log B} \right) \tag{2}
\]
With some additional manipulation and properties of the gamma function, we are left with:

\[
F'_B(b) = 1 + 2 \sum_{m=1}^{M-1} \text{Re} \left[ e^{-2\pi im \left( b - \frac{\log \alpha}{\log B} \right) \cdot \Gamma \left( 1 + \frac{2\pi im}{\gamma \log B} \right)} \right] \\
+ 2 \sum_{m=M}^{\infty} \left[ e^{-2\pi im \left( b - \frac{\log \alpha}{\log B} \right) \cdot \Gamma \left( 1 + \frac{2\pi im}{\gamma \log B} \right)} \right].
\]

Acknowledgements

This work was primarily done at Williams College through the SMALL REU program, under the direction of Dr. Steven J. Miller. The project is supported by NSF grants DMS0850577 and DMS0970067.