Theory and Applications of Benford's Law

Victoria Cuff and Allison Lewis; Advisor: Steven J. Miller

Number Theory and Probability Group - SMALL 2010 - Williams College

1. Background

Definition 1.1. *Benford's Law of Leading Digit Bias states that in many real-life data sets, the proportion of values beginning with digit d is $\log_B(1+\frac{1}{d})$.*

The Benford distribution of leading digits base 10 is:

Leading Digit	Benford Base 10 Probability
1	0.30103
2	0.17609
3	0.12494
4	0.09691
5	0.07918
6	0.06695
7	0.05799
8	0.05115
9	0.04576

Why do we care about Benford's Law?
- ◇ Fraud detection and data integrity
- ◇ Errors in rounding or data collection methods

Benford Tests: In the applications section, we utilize a variety of Benford tests, including comparing the beginning digits to the Benford probabilities, testing the last digit against the uniform distribution, and comparing the proportion of occurrences of combinations of the final two digits relative to one another.

Potential Issues
- ◇ Chi-square statistic is extremely sensitive to large data sets - absolute mean deviation is often a better measure of conformity.
- ◇ Possibility that the data sets are not *supposed* to follow the Benford distribution.

Abstract

In this study, we demonstrate applications of Benford's Law in the analysis of several diverse data sets, including voting records from the 2009 Iranian election and a portion of the data from the Climategate scandal. We analyze each data set for conformity to Benford's Law, and consider possible instances of rounding discrepancies, errors in data collection methods, or fraud. We finish by looking at the theoretical aspect of Benford's Law, determining how closely various Weibull distributions follow the Benford distribution and analyzing how varying the parameters affects the Weibull's conformance to the expected leading digit probabilities.

2. Applications of Benford's Law

2.1 Iranian Election 2009

- Controversial presidential election in 2009
- Suspicion of ballot-stuffing fraud
- Polling vs. Precinct
  - ◇ Polling: over 45,000 observations per each candidate
  - ◇ Precinct: 320 observations per candidate

Test	Total	Ahmadinejad	Mousavi	95%
First Digit	29.14	36.84	9.92	15.5
Last Digit	11.24	8.71	9.10	16.9
Endings	114.88	99.93	102.17	124.3
Non/Doubles	3.47	0.99	1.03	3.8
Non/Doubles(S)	27.74	10.23	10.53	16.9
Doubles(C)	18.82	9.13	9.33	15.5

Table 1: *Chi-Square Means: Polling Level (Split)*

Conclusions Possible explanations for the significant deviations include higher voter turnout (from a previously silent majority) or growth in support for Ahmadinejad. However, these explanations cannot account for a voter turnout of $\geq 100\%$ for two provinces. The significant results for Ahmadinejad suggest the possibility of ballot-box stuffing.

2.2 Climate Data

- Thousands of CRU emails leaked in November 2009
- Allegations of scientific misconduct in the climate science community

Problem: Amalgamation of all thirty data subsets gave spike of values ending in 77 and deficit of values ending in 00.
Approach: Analyze subsets of data with strange last two digit distributions:

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

Data Set	00	11	22	33	44	55	66	77	88	99
West. US	4	6	4	5	1	8	0	38	0	24
Tasmania	57	80	64	57	0	0	0	0	0	0

Table 2: *Ending Double-Digit Occurrences in Select Data Series*

"Tasmania Unsmoothed" Analysis: 46 ending combinations unobserved.

Test	Chi-Square	Abs. Mean Dev.
Endings	3261.49	1.13
Non/Doubles	19.36	2.96
Non/Doubles(S)	538.58	1.63
Doubles(C)	400.68	12.00

Table 3: *"Tasmania Unsmoothed" Data: Last Two Digits Tests*

Conclusions Similar analysis on all 30 data subsets reveals multiple cases of suspicious disparities from Benford. These results could be indicative of fraud / data manipulation in the climate data, or could be due to other factors (rounding discrepancies, data collection methods, or non-Benford behavior).

3. Theory of Benford's Law

Question: How close does the distribution of digits of a random variable with a Weibull distribution follow Benford's Law? As we vary the parameters, how does this effect the Weibull distribution's conformance to the expected leading digit probabilities?

Weibull Distribution

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

$$x \geq \beta;\ \gamma,\alpha > 0$$

Poisson Summation and Fourier Transformation: As long as the function is rapidly decaying, we may apply the Fourier Transform, thus

$$H : \widehat{H}(u) = \int_{-\infty}^{\infty} H(t)e^{-2\pi itu}dt.$$

where $\widehat{H}$ is the Poisson Summation of

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

Converting a long, slowly converging sum to a short rapidly converging sum. Thus allowing us to evaluate fewer terms and still achieving accuracy.

Proof:
Let ζ be a Weibull distribution with $\beta = 0$ and $[a,b] \subset [0,1]$.

$$\begin{aligned} F_B(b) &:= \text{Prob}(\log_B \zeta \bmod 1 \in [0,b]) \\ &= \textstyle\sum_{k=-\infty}^{\infty} \text{Prob}(\log_B \zeta \in [0+k, b+k]) \\ &= \text{Prob}(\log_B \zeta \bmod 1 \in [0,b]) \\ &= \textstyle\sum_{k=-\infty}^{\infty} \left(e^{-\left(\frac{B^k}{\alpha}\right)^{\gamma}} - e^{-\left(\frac{B^{b+k}}{\alpha}\right)^{\gamma}}\right) \end{aligned}$$

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

where for $b \in [0,1]$, let $Z = B^b$.

$$F'_B(b) = \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\alpha}\cdot e^{-\left(\frac{ZB^k}{\alpha}\right)^{\gamma}} ZB^{k}\left(\frac{ZB^{k}}{\alpha}\right)^{\gamma-1}\gamma\log B\cdot e^{-2\pi itk}dt$$

With some manipulation and the Gamma function (and its properties) we are left with:

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

Figure 1: *Kolmogorov-Smirnov Test: Comparing the cumulative distribution function of the Weibull Distribution and the Uniform Distribution, when equal (ideal) it is zero.*

4. Acknowledgements

We wish to thank Williams College, whose generous support made this research possible. The authors are funded by NSF grant DMS0850577, and the advisor is funded by NSF grant DMS0970067.

References

[1] Steven J. Miller, Mark J. Nigrini, *Benford's Law Applied to Hydrology Data - Results and Relevance to Other Geophysical Data*, Mathematical Geology, (39 (2007), no. 5, 469–490)..

[2] P.D. Jones, M.E. Mann,*Climate Over Past Millennia*, 2004, IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series # 2004-085, NOAA/NGDC Paleoclimatology Program, Boulder CO, USA.

[3] F. Benford, *The Law of Anomalous Numbers*, Proceedings of the American Philosophical Society, Vol. 78.

[4] Mark J. Nigrini, *A Taxpayer Compliance Application of Benford's Law*, J Am Tax Assoc 18:72-91, 1996.

[5] Walter Mebane, *Note on the Presidential Election in Iran*, 2009.

[6] Ali Ansari, Daniel Berman, and Thomas Rintoul, *Preliminary Analysis of the Voting Figures in Iran's 2009 Presidential Election*, Chatham House: Independent thinking on international affairs, 2009.