Theory and Applications of Benford's Law

Victoria Cuff
Clemson University

Allison Lewis
University of Portland

Advisor: Steven J. Miller

SMALL 2010 - Williams College

Summary

- Review Benford's Law
- Applications of Benford's Law:
  - ◇ Iranian Election Results of 2009
  - ◇ Climategate Data
- Theory of Benford's Law:
  - ◇ IRS Project
  - ◇ Weibull Distribution

Benford's Law: Newcomb (1881), Benford (1938)

Statement

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

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

Benford Tests

First and Last Digit Tests:

- First Digit
  ◇ $P(d_1) = \log_B(1 + \frac{1}{d_1})$
- First Two Digits
  ◇ $P(d_1 d_2) = \log_B(1 + \frac{1}{10d_1 + d_2})$
- First Three Digits
  ◇ $P(d_1 d_2 d_3) = \log(1 + \frac{1}{100d_1 + 10d_2 + d_3})$
- Last Digit
  ◇ P(last digit d)$= \frac{1}{10}$

Benford Tests (continued)

Last Two-Digit Tests:

- All Endings
  - ◇ P(any ending $d_1 d_2$)$= \frac{1}{100}$
- Non-Doubles vs. Doubles
  - ◇ P(non-double)$= \frac{9}{10}$, P(double)$= \frac{1}{10}$
- Non-Doubles vs. Doubles (Split)
  - ◇ P(non-double)$= \frac{9}{10}$, P(any double $d_1 d_1$)$= \frac{1}{100}$
- Doubles (Conditional)
  - ◇ P($d_1 d_1$|double)$= \frac{1}{10}$

Note: Chi-square statistic is extremely sensitive to large data sets - absolute mean deviation is often a better measure of conformity.

2009 Iranian Election

- Controversial presidential election in 2009
- Suspicion of ballot-stuffing fraud
- Prior Benford Tests:
  - ⋄ Walter Mebane (2009) - Second Digit Analysis
- Data analyzed provided by Mebane
- Polling vs. Precinct
  - ⋄ Polling: over 45,000 observations per each candidate
  - ⋄ Precinct: 320 observations per candidate

Chi-Square Statistics: Polling Level (Split)

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: Chi-Square Means: Polling Level (Split)

Climategate Scandal

- Thousands of CRU emails leaked in November 2009
- Allegations of scientific misconduct in the climate science community
- Refusal to meet FOI Act and release data led to accusations of data distortion

Data Analyzed

- “Proxy Temperature Reconstruction" data from “Global Surface Temperatures Over the Past Two Millenia" (Phil D. Jones, Michael E. Mann)
- Subset of data containing 32,451 observations - further split into 30 data subsets covering data in different regions of the world

Last Two Digit Analysis

Amalgamation of all thirty data subsets gave spike of values ending in 77 and deficit of values ending in 00:

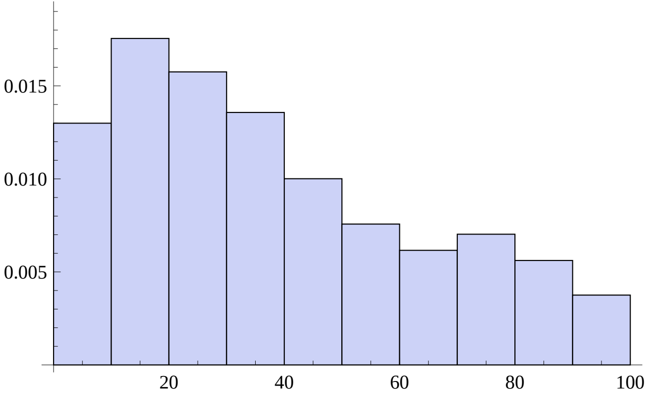

Figure: Double-digit ending combinations in climate data

Approach

Analyze subsets of data with strange last two digit distributions:

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

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: Ending Double-Digit Occurrences in Select Data Series

“Tasmania" Analysis

- 46 ending combinations not observed at all
- Range: [-4.43, 3.59]

00	01	02	03	04	05	06	07	08	09	10	11
57	0	0	72	2	0	79	0	49	2	0	80

Table: First 12 Ending Digit Occurrences for Tasmania Unsmoothed

“Tasmania" Analysis (continued)

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: “Tasmania Unsmoothed" Data: Last Two Digits Tests

Climate Data Conclusions

Conclusion

A similar analysis can be performed on all thirty data subsets, revealing multiple cases of suspicious disparities from the Uniform distribution. These strange results could be indicative of instances of fraud and data manipulation contained in the climate data, or could possibly be due to other factors such as rounding discrepancies and data collection methods.

IRS Project

IRS Project

- CONFIDENTIAL!

Weibull Distribution

Weibull Distribution

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

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

- 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?

Fourier Transform

As long as the function is rapidly decaying, we may apply the Fourier Transform, thus

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

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 $[a, b] \subset [0, 1]$.

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

Proof

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

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

Proof

$$F'_B(b) = \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\alpha} \cdot e^{-\left(\frac{ZB^k}{\alpha}\right)^{\gamma}} ZB^k \left(\frac{ZB^k}{\alpha}\right)^{\gamma-1} \gamma \log B \cdot e^{-2\pi i t k} dt$$

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

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

Kolmogorov-Smirnov Test

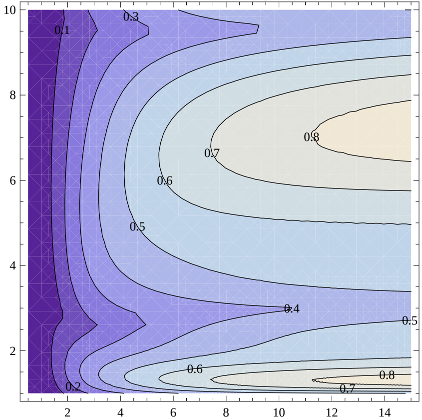

Figure: K-S Test: Comparing the cumulative distribution function of the Weibull Distribution and the Uniform Distribution, when equal (ideal) it is zero.

Kolmogorov-Smirnov Test

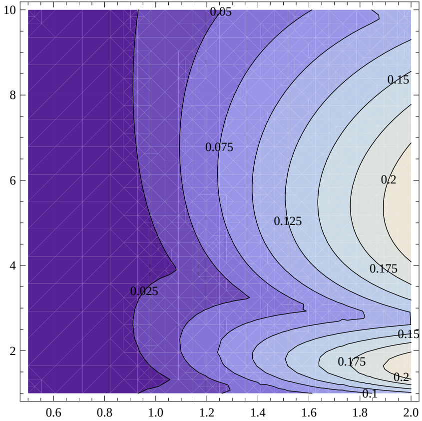

Figure: K-S Test: Comparing the cumulative distribution function of the Weibull Distribution and the Uniform Distribution, when equal (ideal) it is zero.