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

Allison Lewis
University of Portland

NUMS Conference
April 9th, 2011

Summary

- **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

Notation

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: Newcomb (1881), Benford (1938)

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 $\log_B s$.

Benford Base 10 Probabilities

For a Benford base 10 data set, we expect the leading digits to (approximately) follow the proportions below:

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 Base 10 Probabilities - First Four Digits

We can extend these proportions as far out into the mantissa as we wish:

	Position in Number			
Digit	**1st**	**2nd**	**3rd**	**4th**
0		0.11968	0.10178	0.10018
1	0.30103	0.11389	0.10138	0.10014
2	0.17609	0.10882	0.10097	0.10010
3	0.12494	0.10433	0.10057	0.10006
4	0.09691	0.10031	0.10018	0.10002
5	0.07918	0.09668	0.09979	0.09998
6	0.06695	0.09337	0.09940	0.09994
7	0.05799	0.09035	0.09902	0.09990
8	0.05115	0.08757	0.09864	0.09986
9	0.04576	0.08500	0.09827	0.09982

So What?

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.

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)$= \frac{1}{10}$

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

Disclaimer

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

First Three Digits Test

In this study, we analyze the first three digits.

Recall, the probability of observing a mantissa that begins with $d_1 d_2 d_3$ is:

$$\log_{10}\left(1 + \frac{1}{100d_1 + 10d_2 + d_3}\right)$$

Restricting the Data Set

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

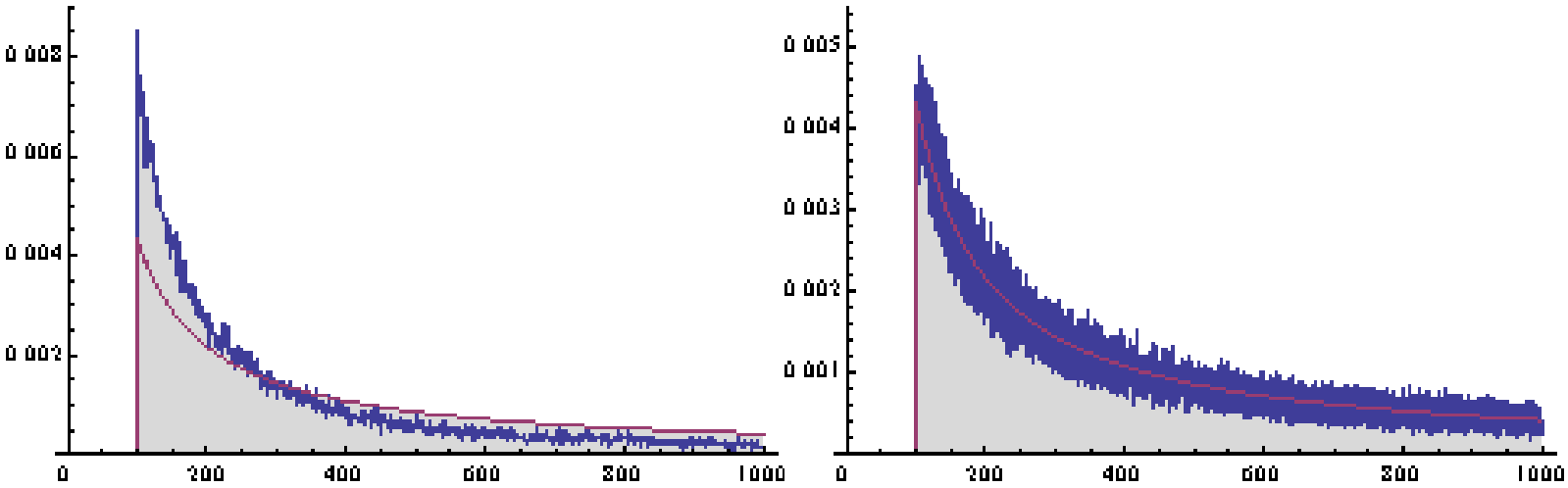

Figure: (Left: Restricted; Right: Unrestricted) Comparing the First Three Digits Tests

Chi-Square and MAD Analysis

Test	Chi-Square	MAD
First Digit	45.82	0.00086
First 2 Digits	178.74	0.00017
First 3 Digits (Restricted)	12054.70	0.00039
First 3 Digits (Unrestricted)	23345.30	0.00020

Table: Starting Digit Tests: Hydrology Data

Comparing Benford Characteristics

Characteristic	Restricted	Unrestricted
Size of Data Set	73,828	446,055
Orders of Magnitude	1	6
# Significant Digits	≥ 4	≥ 3
Mean Absolute Deviation	0.00039	0.00020

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

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

Table: Polling Level - 45,692 observations

Precinct Level Statistics

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

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)

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

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

Iranian Election Conclusions

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.

Last Two Digit Analysis

Problem:

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

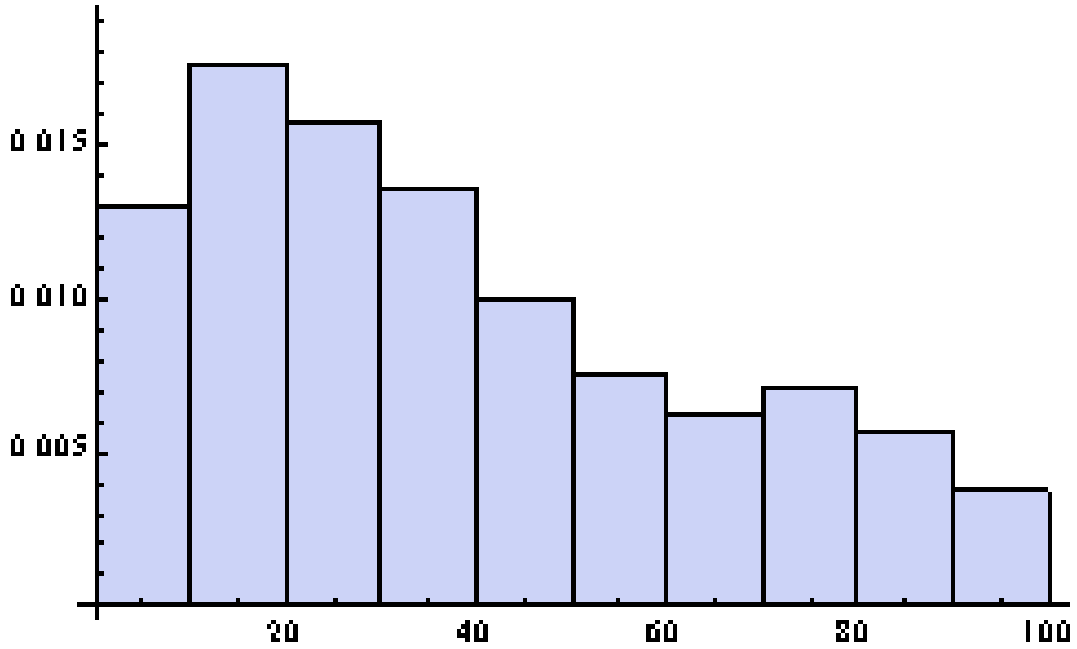

Figure: Double-digit ending combinations in climate data

Approach

Analyze subsets of data with strange last two digit distributions:

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

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	Mean Abs. Dev.
Endings	3261.49	0.0113
Non/Doubles	19.36	0.0296
Non/Doubles(S)	538.58	0.0163
Doubles(C)	400.68	0.1200

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

Climate Data Conclusions

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

Theoretical Questions

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

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

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

- 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

Statement

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

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

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

$$F_B'(z) = 1 + 2\sum_{m=1}^{M-1} \text{Re}\left[e^{-2\pi i m\left(z-\frac{\log\alpha}{\log B}\right)} \cdot \Gamma\left(1+\frac{2\pi i m}{\gamma \log B}\right)\right] + \mathcal{E}\left(\frac{2\sqrt{2}M}{\pi^3}(40+\pi^2)\sqrt{\gamma \log B}\cdot e^{-\pi^2 M/\gamma \log B}\right)$$

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 ϵ 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

$$k = -\ln\left(\frac{a\epsilon}{C}\right), \quad a = \frac{\pi^2}{\gamma \log B}, \quad C = \frac{2\sqrt{2}(40 + \pi^2)\sqrt{\gamma \log B}}{\pi^3}.$$

Kolmogorov - Smirnov Test

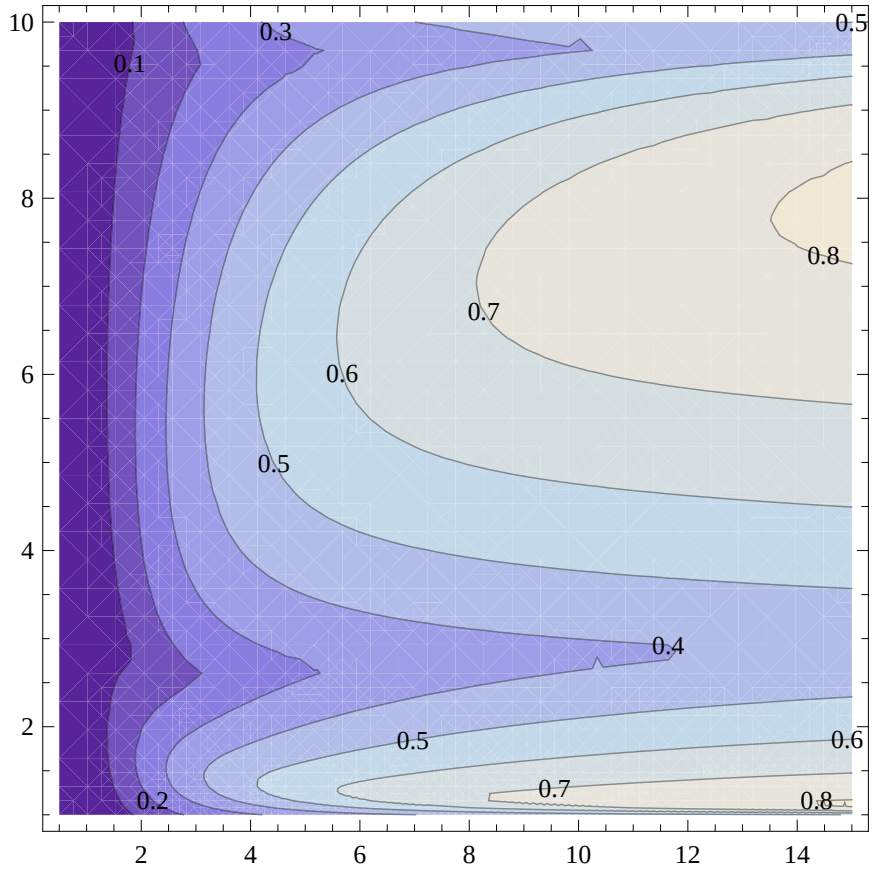

Figure: $\gamma \in [0, 15]$. As γ increases, the Weibull distribution is no longer a good fit compared to the uniform. Note that α has less of an effect on the overall conformance.

Kolmogorov - Smirnov Test

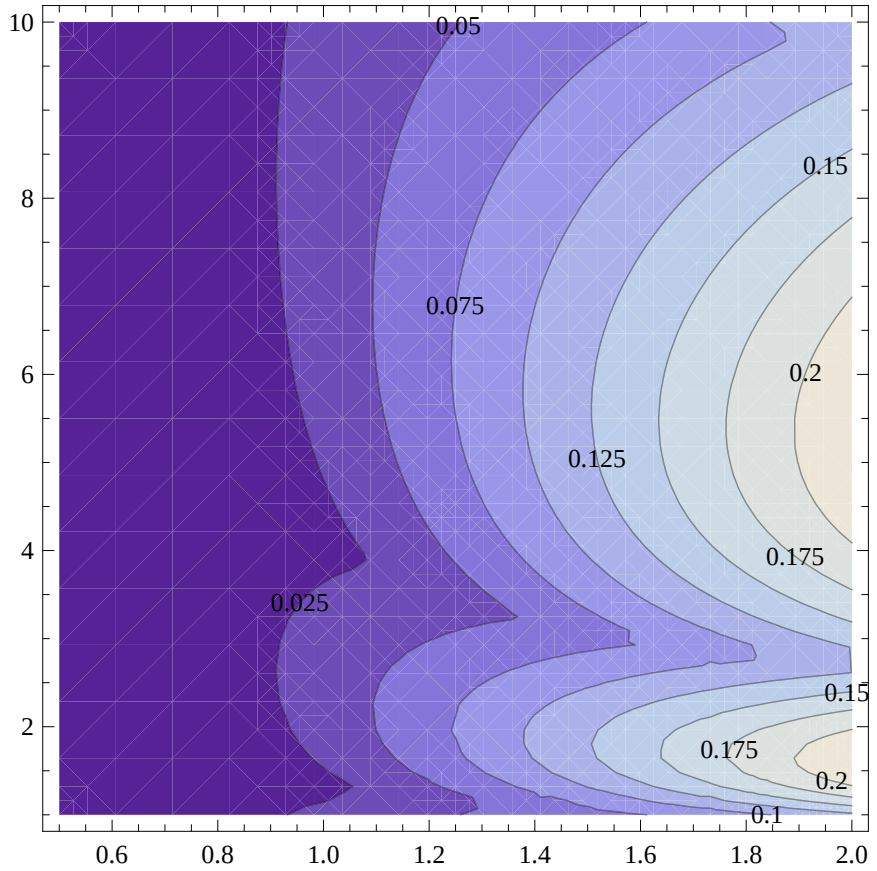

Figure: $\gamma \in [0, 2]$. As γ increases, the Weibull distribution is no longer a good fit compared to the uniform. Note that α has less of an effect on the overall conformance.

Approach

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.

Poisson Summation and Fourier Transform

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

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

where $\widehat{H}$ is the Fourier Transform of

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

Proof of Theorem (Part 1)

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

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

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:

$$\begin{aligned} F'_B(z) &= \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-w} \cdot \exp\left(-2\pi i k \cdot \log_B\left(\frac{\alpha w^{1/\gamma}}{\zeta}\right)\right) dw \\ &= \sum_{k=-\infty}^{\infty} \left(\frac{\alpha}{\zeta}\right)^{-2\pi i k/\log B} \int_{-\infty}^{\infty} e^{-w} \cdot w^{-2\pi i k/\gamma \log B} dw \\ &= \sum_{k=-\infty}^{\infty} \left(\frac{\alpha}{\zeta}\right)^{-2\pi i k/\log B} \Gamma\left(1 - \frac{2\pi i k}{\gamma \log B}\right) \end{aligned} \tag{2}$$

Proof of Theorem (Part 1 - continued)

With some additional manipulation and properties of the gamma function, we are left with:

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

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.