# Number Theory and Probability

Nadine Amersi, Thealexa Becker, Olivia Beckwith,
Alec Greaves-Tunnell, Geoffrey Iyer, Oleg Lazarev,
Ryan Ronan, Karen Shen, Liyang Zhang

Advisor Steven Miller (sjm1@williams.edu)

http://web.williams.edu/Mathematics/sjmiller/public_html/math/talks/talks.html

**Williams College, August 2, 2011**

Generalized Ramanujan Primes
Nadine Amersi, Olivia Beckwith, Ryan Ronan

## Historical Introduction

**Bertrand's Postulate (1845)**

For all integers $x \geq 2$, there exists at least one prime in $(x/2, x]$.

## Ramanujan Primes

**Definition**

The $n$-th Ramanujan prime is the integer $R_n$ that is the smallest to guarantee there are $n$ primes in $(x/2, x]$ for all $x \geq R_n$.

## Ramanujan Primes

**Definition**

The *n*-th Ramanujan prime is the integer $R_n$ that is the smallest to guarantee there are *n* primes in $(x/2, x]$ for all $x \geq R_n$.

**Theorem**

- Ramanujan: For each integer *n*, $R_n$ exists.

## Ramanujan Primes

**Definition**

The $n$-th Ramanujan prime is the integer $R_n$ that is the smallest to guarantee there are $n$ primes in $(x/2, x]$ for all $x \geq R_n$.

**Theorem**

- Ramanujan: For each integer $n$, $R_n$ exists.
- Sondow: $R_n \sim p_{2n}$.

## Ramanujan Primes

**Definition**

The $n$-th Ramanujan prime is the integer $R_n$ that is the smallest to guarantee there are $n$ primes in $(x/2, x]$ for all $x \geq R_n$.

**Theorem**

- Ramanujan: For each integer $n$, $R_n$ exists.
- Sondow: $R_n \sim p_{2n}$.
- Sondow: As $n \to \infty$, $\frac{1}{2}$ of primes are Ramanujan.

## $c$-Ramanujan Primes

### Definition

The $n$-th $c$-Ramanujan prime is the integer $R_{c,n}$ that is the smallest to guarantee there are $n$ primes in $(cx, x]$ for all $x \geq R_{c,n}$ where $c \in (0, 1)$.

## $c$-Ramanujan Primes

**Definition**

The $n$-th $c$-Ramanujan prime is the integer $R_{c,n}$ that is the smallest to guarantee there are $n$ primes in $(cx, x]$ for all $x \geq R_{c,n}$ where $c \in (0, 1)$.

- For each $c$ and integer $n$, does $R_{c,n}$ exist?

# $c$-Ramanujan Primes

**Definition**

The $n$-th $c$-Ramanujan prime is the integer $R_{c,n}$ that is the smallest to guarantee there are $n$ primes in $(cx, x]$ for all $x \geq R_{c,n}$ where $c \in (0, 1)$.

- For each $c$ and integer $n$, does $R_{c,n}$ exist? Yes!

## $c$-Ramanujan Primes

**Definition**

The $n$-th $c$-Ramanujan prime is the integer $R_{c,n}$ that is the smallest to guarantee there are $n$ primes in $(cx, x]$ for all $x \geq R_{c,n}$ where $c \in (0, 1)$.

- For each $c$ and integer $n$, does $R_{c,n}$ exist? Yes!
- Does $R_{c,n}$ exhibit asymptotic behavior?

## $c$-Ramanujan Primes

**Definition**

The $n$-th $c$-Ramanujan prime is the integer $R_{c,n}$ that is the smallest to guarantee there are $n$ primes in $(cx, x]$ for all $x \geq R_{c,n}$ where $c \in (0, 1)$.

- For each $c$ and integer $n$, does $R_{c,n}$ exist? Yes!
- Does $R_{c,n}$ exhibit asymptotic behavior? $R_{c,n} \sim p_{\frac{n}{1-c}}$

## $c$-Ramanujan Primes

### Definition

The $n$-th $c$-Ramanujan prime is the integer $R_{c,n}$ that is the smallest to guarantee there are $n$ primes in $(cx, x]$ for all $x \geq R_{c,n}$ where $c \in (0, 1)$.

- For each $c$ and integer $n$, does $R_{c,n}$ exist? Yes!
- Does $R_{c,n}$ exhibit asymptotic behavior? $R_{c,n} \sim p_{\frac{n}{1-c}}$
- As $n \to \infty$, what proportion of primes are $c$-Ramanujan?

## $c$-Ramanujan Primes

**Definition**

The $n$-th $c$-Ramanujan prime is the integer $R_{c,n}$ that is the smallest to guarantee there are $n$ primes in $(cx, x]$ for all $x \geq R_{c,n}$ where $c \in (0, 1)$.

- For each $c$ and integer $n$, does $R_{c,n}$ exist? Yes!
- Does $R_{c,n}$ exhibit asymptotic behavior? $R_{c,n} \sim p_{\frac{n}{1-c}}$
- As $n \to \infty$, what proportion of primes are $c$-Ramanujan? 1-c

## Existence of $R_{c,n}$

**Theorem**

For all $n \in \mathbb{Z}$ and all $c \in (0, 1)$, the $n$-th $c$-Ramanujan prime $R_{c,n}$ exists.

## Existence of $R_{c,n}$

**Theorem**

For all $n \in \mathbb{Z}$ and all $c \in (0, 1)$, the $n$-th $c$-Ramanujan prime $R_{c,n}$ exists.

Sketch of proof:

## Existence of $R_{c,n}$

**Theorem**

For all $n \in \mathbb{Z}$ and all $c \in (0, 1)$, the $n$-th $c$-Ramanujan prime $R_{c,n}$ exists.

Sketch of proof:

- Let $\pi(x)$ denote the number of primes at most $x$. Then the number of primes in $(cx, x]$ is $\pi(x) - \pi(cx)$.

## Existence of $R_{c,n}$

**Theorem**

For all $n \in \mathbb{Z}$ and all $c \in (0, 1)$, the $n$-th $c$-Ramanujan prime $R_{c,n}$ exists.

Sketch of proof:

- Let $\pi(x)$ denote the number of primes at most $x$. Then the number of primes in $(cx, x]$ is $\pi(x) - \pi(cx)$.
- Using the Prime Number Theorem and Mean Value Theorem, we obtain $\pi(x) - \pi(cx) = \frac{(1-c)x}{\log x} + O\left(\frac{x}{\log^2 x}\right)$.

## Existence of $R_{c,n}$

**Theorem**

For all $n \in \mathbb{Z}$ and all $c \in (0, 1)$, the $n$-th $c$-Ramanujan prime $R_{c,n}$ exists.

Sketch of proof:

- Let $\pi(x)$ denote the number of primes at most $x$. Then the number of primes in $(cx, x]$ is $\pi(x) - \pi(cx)$.
- Using the Prime Number Theorem and Mean Value Theorem, we obtain $\pi(x) - \pi(cx) = \frac{(1-c)x}{\log x} + O\left(\frac{x}{\log^2 x}\right)$.
- $\pi(x) - \pi(cx) \geq n$ for all $x$ sufficiently large.

## Distribution of generalized Ramanujan primes

Expected longest run $\approx \log_{1/p}(n(1-p))$.

| | Length of the longest run below $10^6$ of | | | |
|---|---|---|---|---|
| | *c*-Ramanujan primes | | Non-Ramanujan primes | |
| c | Expected | Actual | Expected | Actual |
| 0.05 | 127 | 97 | 4 | 2 |
| 0.10 | 71 | 58 | 5 | 3 |
| 0.15 | 50 | 42 | 6 | 6 |
| 0.20 | 38 | 36 | 8 | 7 |
| 0.25 | 31 | 27 | 9 | 12 |
| 0.30 | 25 | 25 | 10 | 12 |
| 0.35 | 22 | 18 | 11 | 18 |
| 0.40 | 19 | 21 | 13 | 16 |
| 0.45 | 16 | 19 | 15 | 23 |
| 0.50 | 14 | 20 | 17 | 36 |

More Sums Than Differences Sets

Geoff Iyer, Oleg Lazarev, Liyang Zhang

## Statement

$A$ finite set of integers, $|A|$ its size. Form

- Sumset: $A + A = \{a_i + a_j : a_i, a_j \in A\}$.
- Difference set: $A - A = \{a_i - a_j : a_j, a_j \in A\}$.

## Statement

$A$ finite set of integers, $|A|$ its size. Form

- Sumset: $A + A = \{a_i + a_j : a_i, a_j \in A\}$.
- Difference set: $A - A = \{a_i - a_j : a_i, a_j \in A\}$.

### Definition

We say $A$ is difference dominated if $|A - A| > |A + A|$, balanced if $|A - A| = |A + A|$ and sum dominated (or an MSTD set) if $|A + A| > |A - A|$.

## Statement

$A$ finite set of integers, $|A|$ its size. Form

- Sumset: $A + A = \{a_i + a_j : a_i, a_j \in A\}$.
- Difference set: $A - A = \{a_i - a_j : a_i, a_j \in A\}$.

### Definition

We say $A$ is difference dominated if $|A - A| > |A + A|$, balanced if $|A - A| = |A + A|$ and sum dominated (or an MSTD set) if $|A + A| > |A - A|$.

### Definition

$$kA \ = \ \underbrace{A + \dots + A}_{\text{k times}}, \quad [a, b] \ = \ \{a, a+1, \dots, b\}.$$

## Questions

- Can we find a set $A$ such that $|kA + kA| > |kA - kA|$?
- Can we find a set $A$ such that $|A + A| > |A - A|$ and $|2A + 2A| > |2A - 2A|$?
- Can we find a set $A$ such that $|kA + kA| > |kA - kA|$ for all $k$?

## Questions

- Can we find a set $A$ such that $|kA + kA| > |kA - kA|$? YES!
- Can we find a set $A$ such that $|A + A| > |A - A|$ and $|2A + 2A| > |2A - 2A|$? YES!
- Can we find a set $A$ such that $|kA + kA| > |kA - kA|$ for all $k$? NO! (No such set exists)

## $|kA + kA| > |kA - kA|$

Question: Can we find a set $A$ such that $|kA + kA| > |kA - kA|$? YES!

## $|kA + kA| > |kA - kA|$

Question: Can we find a set $A$ such that $|kA + kA| > |kA - kA|$? YES!

Example: $|3A + 3A| > |3A - 3A|$

$$\begin{aligned} A &= [0,12] \cup [16,18] \cup \{24\} \cup [139,161] \\ &\cup \{275\} \cup [281,283] \cup [287,300] \end{aligned}$$

$|3A + 3A| = 1798, |3A - 3A| = 1795.$

## Generalizations

By further modifying $A$, we can construct sets where

- The sumset has arbitrarily more elements than the difference set:
  $|kA + kA| - |kA - kA| = m$
- The sumset and difference set each have arbitrarily many missing elements:
  $|kA + kA| = 2nk + 1 - m$ and $|kA - kA| = 2nk + 1 - \ell$
  for any $m, \ell$ such that $\ell \leq 2m$
- $|s_1A - d_1A| = (s_1 + d_1)n + 1 - m$ and
  $|s_2A - d_2A| = (s_2 + d_2)n + 1 - \ell$
  for $\ell \leq 2m$ and $s_1 + d_1 = s_2 + d_2$

## $k$-Generational Sets

Question: Does a set $A$ exist such that $|A+A| > |A-A|$ and $|A+A+A+A| > |A+A-A-A|$? If yes call it 2-generational.

## *k*-Generational Sets

Question: Does a set $A$ exist such that $|A+A| > |A-A|$ and $|A+A+A+A| > |A+A-A-A|$? If yes call it 2-generational.

Yes!

$$A = \{0, 1, 3, 4, 5, 26, 27, 29, 30, 33, 37, 38, 40, 41, 42, 43, 46, 49, 50, 52, 53, 54, 72, 75, 76, 79, 80\}$$

In fact, we can do much better.

## $k$-Generational Sets

We can find an $A$ such that $|x_jA - y_jA| > |w_jA - z_jA|$ for any nontrivial choices of $x_j, y_j, w_j, z_j$ and for all $2 \leq j \leq k$.

## $k$-Generational Sets

We can find an $A$ such that $|x_jA - y_jA| > |w_jA - z_jA|$ for any nontrivial choices of $x_j, y_j, w_j, z_j$ and for all $2 \leq j \leq k$.

Example: We can find an $A$ such that

$$
\begin{gathered}
|A + A| > |A - A| \\
|A + A - A| > |A + A + A| \\
|5A - 2A| > |A - 6A| \\
\vdots \\
|1870A - 141A| > |1817A - 194A|
\end{gathered}
$$

Base Expansion: For sets $A_1, \ldots, A_n$ and $m \in \mathbb{N}$ sufficiently large (relative to $k$ and $A_1, \ldots, A_n$) the set

$$A = A_1 + m \cdot A_2 + \cdots + m^{n-1} \cdot A_n$$

(where the multiplication is the usual scalar multiplication) has

$$|xA - yA| = \prod_{j=1}^{k} |xA_j - yA_j|$$

whenever $x + y \leq k$.

Base Expansion: For sets $A_1, \ldots, A_n$ and $m \in \mathbb{N}$ sufficiently large (relative to $k$ and $A_1, \ldots, A_n$) the set

$$A = A_1 + m \cdot A_2 + \cdots + m^{n-1} \cdot A_n$$

(where the multiplication is the usual scalar multiplication) has

$$|xA - yA| = \prod_{j=1}^{k} |xA_j - yA_j|$$

whenever $x + y \leq k$.

Base expansion is an approximation to the cross product. However, it only works for finitely many sums/differences.

## $k$-Generational Sets

To prove the theorem, we choose sets $A_j$ that behave well for a specific $2 \leq j \leq k$ and are balanced for $i \neq j$. We then use base expansion to create $A$ using the $A_j$.

We can construct $k$-generation sets for arbitrarily large $k$. But for any set $A$, as $k$ goes to infinity $kA$ will become difference-dominated or balanced.

We can construct $k$-generation sets for arbitrarily large $k$. But for any set $A$, as $k$ goes to infinity $kA$ will become difference-dominated or balanced.

**Theorem (Nathanson)**

For any set $A$, as $k$ goes to infinity the fringes of $kA$ will stabilize. If the largest element of $A$ is $a$ and there are $m$ elements in $A$, $kA$ will stabilize before $k = a^2 m$.

We can construct $k$-generation sets for arbitrarily large $k$. But for any set $A$, as $k$ goes to infinity $kA$ will become difference-dominated or balanced.

**Theorem (Nathanson)**

For any set $A$, as $k$ goes to infinity the fringes of $kA$ will stabilize. If the largest element of $A$ is $a$ and there are $m$ elements in $A$, $kA$ will stabilize before $k = a^2m$.

Here we will improve this bound.

**Theorem**

For any set $A$, as $k$ goes to infinity the fringes of $kA$ will stabilize. If the largest element of $A$ is $a$ and there are $m$ elements in $A$, $kA$ will stabilize before $k = a$.

**Theorem**

For any set $A$, as $k$ goes to infinity the fringes of $kA$ will stabilize. If the largest element of $A$ is $a$ and there are $m$ elements in $A$, $kA$ will stabilize before $k = a$.

**Theorem**

For any set $A$, as $k$ goes to infinity $kA$ will eventually become difference-dominated or balanced. And this will happen before $k$ reaches $2a$.

Proof Idea: $kA \subset kA - kA$. And $k(A - A)$ and $2k(A)$ will both become stabilize when $k = 2a$.

Random Matrix Theory
Olivia Beckwith, Karen Shen

## Random Matrices

Distribution of eigenvalues of random matrices: $Ax = \lambda x$.

## Random Matrices

Distribution of eigenvalues of random matrices: $Ax = \lambda x$.

Applications:

## Random Matrices

Distribution of eigenvalues of random matrices: $Ax = \lambda x$.

Applications:

- Nuclear Physics

## Random Matrices

Distribution of eigenvalues of random matrices: $Ax = \lambda x$.

Applications:

- Nuclear Physics
- L-functions

## Matrix Ensembles

$$\text{Toeplitz: } \begin{pmatrix} b_0 & b_1 & b_2 & b_3 \\ b_1 & b_0 & b_1 & b_2 \\ b_2 & b_1 & b_0 & b_1 \\ b_3 & b_2 & b_1 & b_0 \end{pmatrix}$$

## Matrix Ensembles

Toeplitz: $\begin{pmatrix} b_0 & b_1 & b_2 & b_3 \\ b_1 & b_0 & b_1 & b_2 \\ b_2 & b_1 & b_0 & b_1 \\ b_3 & b_2 & b_1 & b_0 \end{pmatrix}$

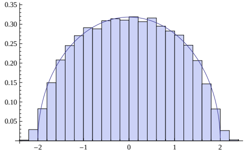
Signed Toeplitz:

$$\begin{pmatrix} b_0 & -b_1 & b_2 & b_3 \\ -b_1 & -b_0 & b_1 & -b_2 \\ b_2 & b_1 & b_0 & -b_1 \\ b_3 & -b_2 & -b_1 & b_0 \end{pmatrix}$$

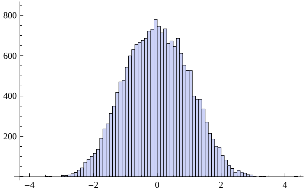
$a_{ii} = \epsilon_{ii} a = \pm a$, $p = \text{Prob}(\epsilon_{ij} = 1), \dots$

# Previous Work

Real symmetric:




Toeplitz:



## Methods: Markov's Method of Moments

- The $k^{\text{th}}$ moment $M_k$ of a probability distribution $f(x)$ defined on an interval $[a, b]$ is $\int_a^b x^k f(x)\, dx$.

## Methods: Markov's Method of Moments

- The $k^{\text{th}}$ moment $M_k$ of a probability distribution $f(x)$ defined on an interval $[a, b]$ is $\int_a^b x^k f(x)\, dx$.
- Show a typical eigenvalue measure $\mu_{A,N}(x)$ converges to a probability distribution $P$ by controlling convergence of average moments of the measures as $N \to \infty$ to the moments of $P$.

# Eigenvalue Trace Lemma

For any non-negative integer $k$, if $A$ is an $N \times N$ matrix with eigenvalues $\lambda_i(A)$, then

$$Trace\left(A^k\right) = \sum_{i=1}^{N} \lambda_i(A)^k.$$

## Eigenvalue Trace Lemma

For any non-negative integer $k$, if $A$ is an $N \times N$ matrix with eigenvalues $\lambda_i(A)$, then

$$Trace\left(A^k\right) = \sum_{i=1}^{N} \lambda_i(A)^k .$$

Using this lemma, we see that a formula for the average $k^{\text{th}}$ moment, $M_k(N) = \mathbb{E}\left[M_k(A_N)\right]$, is:

$$\frac{1}{N^{\frac{k}{2}+1}} \sum_{1 \le i_1, \dots, i_k \le N} \mathbb{E}\left(\epsilon_{i_1 i_2} b_{|i_1 - i_2|} \epsilon_{i_2 i_3} b_{|i_2 - i_3|} \dots \epsilon_{i_k i_1} b_{|i_k - i_1|}\right)$$

## Evaluating the Moments

$$M_k(N) = \frac{1}{N^{\frac{k}{2}+1}} \sum_{1 \le i_1, \dots, i_k \le N} \mathbb{E}\left(\epsilon_{i_1 i_2} b_{|i_1 - i_2|} \epsilon_{i_2 i_3} b_{|i_2 - i_3|} \dots \epsilon_{i_k i_1} b_{|i_k - i_1|}\right)$$

## Evaluating the Moments

$$M_k(N) = \frac{1}{N^{\frac{k}{2}+1}} \sum_{1 \le i_1, \dots, i_k \le N} \mathbb{E}\left(\epsilon_{i_1 i_2} b_{|i_1 - i_2|} \epsilon_{i_2 i_3} b_{|i_2 - i_3|} \cdots \epsilon_{i_k i_1} b_{|i_k - i_1|}\right)$$

For a term to contribute in the summand:

- The $b$'s must be matched in at least pairs since $\mathbb{E}(b_{ij}) = 0$.

## Evaluating the Moments

$$M_k(N) = \frac{1}{N^{\frac{k}{2}+1}} \sum_{1 \le i_1, \dots, i_k \le N} \mathbb{E}\left(\epsilon_{i_1 i_2} b_{|i_1 - i_2|} \epsilon_{i_2 i_3} b_{|i_2 - i_3|} \dots \epsilon_{i_k i_1} b_{|i_k - i_1|}\right)$$

For a term to contribute in the summand:

- The $b$'s must be matched in at least pairs since $\mathbb{E}(b_{ij}) = 0$.
- The $b$'s must be matched in at most pairs since there must be at least $\frac{k}{2} + 1$ degrees of freedom.

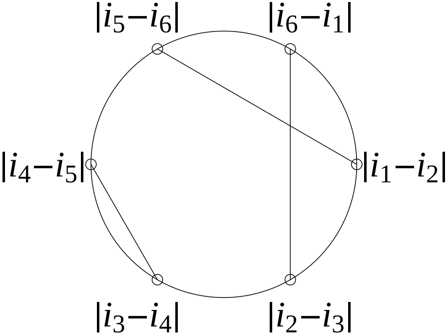
Thus:

Thus:

- Odd moments vanish.

Thus:

- Odd moments vanish.
- For the even moments $M_{2k}$ we can represent each contributing term as a pairing of $2k$ vertices on a circle as follows:



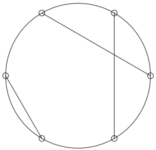
## Results

$p = \frac{1}{2}$: Semi-circle distribution

## Results

$p = \frac{1}{2}$: Semi-circle distribution

$p \neq \frac{1}{2}$: unbounded support.
Each configuration weighted by $(2p-1)^m$, where $m$ is the number of points on the circle whose edge crosses another edge.

Example:  m = 4

## Results, continued

Question: Out of the $(2k-1)!!$ ways to pair $2k$ vertices, how many of these pairings will have $m$ vertices crossing?

## Results, continued

Question: Out of the $(2k-1)!!$ ways to pair $2k$ vertices, how many of these pairings will have $m$ vertices crossing? For:

- $m = 0$, well-known to be the Catalan numbers.

## Results, continued

Question: Out of the $(2k-1)!!$ ways to pair $2k$ vertices, how many of these pairings will have $m$ vertices crossing? For:

- $m = 0$, well-known to be the Catalan numbers.
- $m = 4$, we proved there are $\binom{2k}{k-2}$ such pairings.

## Results, continued

Question: Out of the $(2k-1)!!$ ways to pair $2k$ vertices, how many of these pairings will have $m$ vertices crossing? For:

- $m = 0$, well-known to be the Catalan numbers.
- $m = 4$, we proved there are $\binom{2k}{k-2}$ such pairings.
- $m = 6$, we proved there are $4\binom{2k}{k-3}$ such pairings.

## Results, continued

Question: Out of the $(2k-1)!!$ ways to pair $2k$ vertices, how many of these pairings will have $m$ vertices crossing? For:

- $m = 0$, well-known to be the Catalan numbers.
- $m = 4$, we proved there are $\binom{2k}{k-2}$ such pairings.
- $m = 6$, we proved there are $4\binom{2k}{k-3}$ such pairings.

As $k$ gets very large, the expected number of vertices in a crossing converges to $2k-2$ and the variance converges to 4.

Benford's Law

Thealexa Becker, Alec Greaves-Tunnell, Ryan Ronan

## Benford's Law Review

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

A set is Benford if probability first digit is $d$ is $\log_B\left(\frac{d+1}{d}\right)$; 30% start with 1.

- Many data sets exhibit Benford behavior:
  - ◇ Fibonacci Sequence
  - ◇ Lots of financial data (stocks, bonds, etc.)
  - ◇ Certain products of random independent variables

## Benford's Law Review

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

A set is Benford if probability first digit is $d$ is $\log_B\left(\frac{d+1}{d}\right)$; 30% start with 1.

- Many data sets exhibit Benford behavior:
  - ◇ Fibonacci Sequence
  - ◇ Lots of financial data (stocks, bonds, etc.)
  - ◇ Certain products of random independent variables

**Interesting Question**

Why do we observe Benford distribution of first digits in "real world" data sets?

## Overview

**Lemons' Interesting Answer (American Journal of Physics, 1986)**

Often due to observing distribution of pieces of a conserved quantity.

## Overview

**Lemons' Interesting Answer (American Journal of Physics, 1986)**

Often due to observing distribution of pieces of a conserved quantity.

- Probability model in paper vague and unclear.

## Overview

**Lemons' Interesting Answer (American Journal of Physics, 1986)**

Often due to observing distribution of pieces of a conserved quantity.

- Probability model in paper vague and unclear.

**Proposed model**

Partition $X$ into $N$ terms: $X = \sum_{j=1}^{N} n_j x_j$. Issues: what possible $x_j$'s? Is $N$ fixed?

## Results

- For (small) finite $N$, brute force calculation shows $\mathbb{E}(n_j) = \frac{1}{x_j}(\frac{X}{N})$; Benford density is proportional to $1/x$.

## Results

- For (small) finite $N$, brute force calculation shows $\mathbb{E}(n_j) = \frac{1}{x_j}(\frac{X}{N})$; Benford density is proportional to $1/x$.
- For general $N$, approximate: $S = X - \sum_j n_j x_j$,

$$\delta(X, \sum_{j=1}^{N} n_j x_j) \propto e^{-S^2/2\sigma},$$

then evaluate $N$-dimensional integral.

## Results

- For (small) finite $N$, brute force calculation shows $\mathbb{E}(n_j) = \frac{1}{x_j}(\frac{X}{N})$; Benford density is proportional to $1/x$.
- For general $N$, approximate: $S = X - \sum_j n_j x_j$,

$$\delta(X, \sum_{j=1}^{N} n_j x_j) \ltimes e^{-S^2/2\sigma},$$

then evaluate $N$-dimensional integral.

- Difficulty: region of integration; can simplify with indicator functions, but Fourier transform has slow decay.

## Another model

Consider $M$ sticks of lengths $\ell_i$, each $l_i$ drawn from the random variable $L$. Break each $\ell_i$ by cutting at $k_i \ell_i$, with $K_i \sim \mathrm{Unif}(0, 1)$. Repeat cutting $N$ times.

## Another model

Consider $M$ sticks of lengths $\ell_i$, each $l_i$ drawn from the random variable $L$. Break each $\ell_i$ by cutting at $k_i\ell_i$, with $K_i \sim \mathrm{Unif}(0,1)$. Repeat cutting $N$ times.

## Theorem

If $L$ is Benford on $[1,10)$ and $N = 1$, then as $M \to \infty$ the distribution of lengths of pieces is Benford's Law.

## Another model

Consider $M$ sticks of lengths $\ell_i$, each $l_i$ drawn from the random variable $L$. Break each $\ell_i$ by cutting at $k_i\ell_i$, with $K_i \sim \mathrm{Unif}(0,1)$. Repeat cutting $N$ times.

## Theorem

If $L$ is Benford on $[1,10)$ and $N = 1$, then as $M \to \infty$ the distribution of lengths of pieces is Benford's Law.

◇ Find cumulative probability distribution function of random variable $Z = KL$.

◇ Evaluate

$$\text{Prob[First digit} = d] = \sum_{r=0}^{+\infty} \left[F_z((d+1)10^{-r}) - F_z(d10^{-r})\right].$$

## Another model

Consider $M$ sticks of lengths $\ell_i$, each $l_i$ drawn from the random variable $L$. Break each $\ell_i$ by cutting at $k_i\ell_i$, with $K_i \sim \mathrm{Unif}(0,1)$. Repeat cutting $N$ times.

## Theorem

If $L$ is Benford on $[1,10)$ and $N = 1$, then as $M \to \infty$ the distribution of lengths of pieces is Benford's Law.

◇ Find cumulative probability distribution function of random variable $Z = KL$.

◇ Evaluate

$$\text{Prob[First digit} = d] = \sum_{r=0}^{+\infty} \left[F_z((d+1)10^{-r}) - F_z(d10^{-r})\right].$$

Also true if $N \to \infty$.

**CONJECTURE**

Let $L$ be fixed and consider one stick ($M = 1$). As $N \to \infty$, the resulting first digit distribution of the lengths of the broken pieces will conform to Benford's Law.

$\diamond$ Wish to show that for any digit $d$ the resulting first digit distribution has zero variance.

**CONJECTURE**

Let $L$ be fixed and consider one stick ($M = 1$). As $N \to \infty$, the resulting first digit distribution of the lengths of the broken pieces will conform to Benford's Law.

◇ Wish to show that for any digit $d$ the resulting first digit distribution has zero variance.

◇ Cross terms are most problematic: Need $N \to \infty$ limit of

$$\sum_{i,j=1}^{\infty} \int_{x=0}^{1} \int_{y=0}^{1} \int_{z=\min\left(\frac{10^{-i}}{xy},1\right)}^{\min\left(\frac{2\times 10^{-i}}{xy},1\right)} \int_{w=\min\left(\frac{10^{-j}}{x(1-y)},1\right)}^{\min\left(\frac{2\times 10^{-j}}{x(1-y)},1\right)} \frac{(-\log x)^{n-1}(\log z \log w)^{m-1}}{\Gamma(n)\Gamma(m)^2} dwdzdydx$$

## Definition of Copulas

Copula: A form of joint CDF between multiple variables with given uniform marginals on the d-dimensional unit cube.

## Sklar's Theorem

Let X and Y be random variables with joint distribution function H and marginal distribution functions F and G respectively. There exists a copula, C, such that

$$\forall x, y \in \mathbb{R}, \quad H(x, y) = C(F(x), G(y)).$$

## Archimedean Copulas

A commonly used / studied family of copulas is of the form

$$C(x, y) \;=\; \phi^{-1}(\phi(x) + \phi(y))$$

where $\phi$ is the generator and $\phi^{-1}$ is the inverse generator of the copula.

Investigating the Benfordness of the product of random variables arising from copulas.

Clayton Copula: $C(x,y) = (x^{-\theta} + y^{-\theta} - 1)^{-1/\theta}$.

PDF (bivariate): $\theta(\theta^{-1}+1)(xy)^{-\theta-1}(x^{-\theta}+y^{-\theta}-1)^{-2-1/\theta}$.

PDF (general case):
$$\theta^{n-1}\frac{\Gamma(n+\theta^{-1})}{\Gamma(1+\theta^{-1})}(x_1\cdots x_n)^{-\theta-1}(x_1^{-\theta}+\cdots+x_n^{-\theta}-1)^{-n-1/\theta}.$$

## Results

- Early data and chi-square tests of multivariate copulas suggest Benford behavior of the products of copulas.
- Proof strategy includes the integration of the PDF over the region in which the product has first digit $d$ using Poisson summation:

$$\int_0^1 \cdots \int_0^1 \sum_k \widehat{\phi}_{\log_{10}(x_1 \cdots x_n)}(k) p(x_1, \dots, x_n) dx_1 \cdots dx_n,$$

where

$$\phi_a(u) = \chi_{[1,2)}(10^{u+a}) = \begin{cases} 1 & \text{if } 10^{u+a} \in [1,2) \\ 0 & \text{otherwise.} \end{cases}$$