Benford's Law, MSTD, and c-Ramanujan Primes

Nadine Amersi, Thealexa Becker, Olivia Beckwith, Alexander Greaves-Tunnell, Geoffrey Iyer, Oleg Lazarev, Ryan Ronan, Karen Shen, and Liyang Zhang; Advisor: Steven J. Miller (sjm1@williams.edu)

Number Theory and Probability Group - SMALL 2011 - Williams College

1. Benford's Law

1.1 Introduction

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

Many data sets exhibit Benford behavior: Fibonacci Sequence, Financial data (stocks, bonds, etc.), products of random independent variables

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

1.2 Conserved Quantities

First 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 $1/x$.
- For general N, approximate: $S = N - \sum_j n_j x_j$,

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

  then evaluate N-dimensional integral.
- Error $\leq \sqrt{\frac{2}{\pi}} \frac{1}{k_\sigma e^{k_\sigma^2/2}} \to 0$ as $\sigma \to 0$.

Second Proposed Model Consider M sticks of lengths ℓ_i, each l_i drawn from the random variable L. Break each ℓ_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.

Theorem Let L be some fixed constant such that $l_1 = l_2 = ... = l_M = L$. Then, as $M \to \infty$ and $N \to \infty$, the resulting first digit distribution of the lengths of the broken pieces will conform to Benford's Law.

ConjectureLet 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.

1.3 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 fucntions 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 ϕ is the generator and ϕ^{-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): $\theta^{n-1}\frac{\Gamma(n+\theta^{-1})}{\Gamma(1+\theta^{-1})}(x_1 \cdots x_n)^{-\theta-1}(x_1^{-\theta} + \cdots + x_n^{-\theta} - 1)^{-n-1/\theta}$.

Proof strategy includes the integration of the PDF over the region in which the product has first digit d using Poisson summation.

2. Generalized More-Sum-Than-Difference Sets

2.1 Introduction

A More Sums Than Differences (MSTD, or sum-dominant) set is a finite set $A \subset \mathbb{Z}$ such that $|A + A| < |A - A|$. Though it was believed that the percentage of subsets of $\{0, \ldots, n\}$ that are sum-dominant tends to zero, in 2006 Martin and O'Bryant proved a positive percentage are sum-dominant. We generalize their result to the many different ways of taking sums and differences of a set. We prove that $|\epsilon_1 A + \cdots + \epsilon_k A| > |\delta_1 A + \cdots + \delta_k A|$ a positive percent of the time for all nontrivial choices of $\epsilon_j, \delta_j \in \{-1, 1\}$. Previous approaches proved the existence of infinitely many such sets given the existence of one; however, no method existed to construct such a set. If you are reading this let us know and we'll give you candy. Using base expansion and clever fringe methods, we develop a new, explicit construction for one such set, and then extend to a positive percentage of sets.

We extend these results further, finding sets that exhibit different behavior as more sums/differences are taken. For example, notation as above we prove that for any m, $|\epsilon_1 A + \cdots + \epsilon_k A| - |\delta_1 A + \cdots + \delta_k A| = m$ a positive percentage of the time. We find the limiting behavior of $kA = A + \cdots + A$ for an arbitrary set A as $k \to \infty$ and an upper bound of k for such behavior to settle down. Finally, we say A is k-generational sum-dominant if A, $A + A$, $\ldots$, kA are all sum-dominant. Numerical searches were unable to find even a 2-generational set (heuristics indicate the probability is at most 10^{-9}, and almost surely significantly less). We prove the surprising result that for any k a positive percentage of sets are k-generational, and no set can be k-generational for all k.

2.2 Results

Theorem 2.1. *Given $s_1, d_1, s_2, d_2 \in \mathbb{N} \cup \{0\}$ such that $\{s_1, d_1\} \neq \{s_2, d_2\}$,*

1. There exists a finite set $A \subset \mathbb{Z}$ such that $|s_1 A - d_1 A| > |s_2 A - d_2 A|$.

2. A positive percentage of sets A satisfy the above.

Theorem 2.2 (Chains of Generalized MSTD Sets)**.** *Let x_j, y_j, w_j, z_j be finite sequences of length k such that $x_j + y_j = w_j + z_j = j$, and $\{x_j, y_j\} \neq \{w_j, z_j\}$ for every $2 \leq j \leq k$. A positive percentage of sets A satisfy $\left|x_j A - y_j A\right| > \left|w_j A - z_j A\right|$ for every $2 \leq j \leq k$.*

Theorem 2.3 (Simultaneous Comparisons)**.** *Given finite sequences of length $n \leq \left\lfloor \frac{k}{2} \right\rfloor + 1$ called s_j, d_j such that $s_j + d_j = k$ for all $1 \leq j \leq k$ and $\{s_j, d_j\} \neq \{s_i, d_i\}$ whenever $j \neq i$, there exists a set A such that $|s_n A - d_n A| > \ldots > |s_1 A - d_1 A|$*

Furthermore, we also prove that A can be chosen such that we have an arbitrary difference between $|s_1 A - d_1 A|$ and $|s_2 A - d_2 A|$.

Theorem 2.4 (Arbitrary Differences)**.** *Let a, b, c, d be non-negative integers such that $a > b, c, d$ and $a + b = c + d = q$. If $c \neq d$, then for any non-negative integers m, ℓ such that $\ell \leq 2m$ and all sufficiently large n, exists $A \subseteq [0, n]$ such that $|aA - bA| = qn + 1 - m$ and $|cA - dA| = qn + 1 - \ell$. If $c = d$, then the statement holds with the additional condition that ℓ is even.*

2.3 Examples

Here are some examples of sets that can be produced through the above theorems. If we set

$$A = \{0, 1, 3, 4, 5, 9, 33, 34, 35, 50, 54, 55, 56, 58, 59, 60\}$$

Then $|A + A + A + A| > |A + A + A - A|$. If we have

$$A = \{0, 1, 3, 4, 7, 26, 27, 29, 30, 33, 37, 38, 40, 41, 42, 43, 46, 49, 50, 52, 53, 54, 72, 75, 76, 78, 79, 80\}$$

Then $|A + A| > |A - A|$ and $|A + A + A + A| > |A + A - A - A|$

If we have

$$A = \{0, 1, 3, 4, 5, 6, 11, 50, 51, 53, 54, 55, 56, 61, 97, 132, 137, 138, 140, 142, 143, 144, 182, 187, 188, 189, 190, 192, 193, 194\}$$

Then $|4A - A| > |5A|$ and $|4A - A| > |3A - 2A|$.

3. c-Ramanujan Primes

3.1 Introduction

In 1845, Bertrand conjectured that for all integers x greater than or equal to 2, there exists at least one prime in $(x/2, x]$. This was proved by Chebyshev in 1860, and then generalized by Ramanujan in 1919, who showed for any integer n there is a least prime R_n such that $\pi(x) - \pi(x/2) \geq n$ for all $x \geq R_n$. We generalize the interval of interest by introducing a parameter $c \in (0, 1)$ and defining the nth c-Ramanujan prime $R_{c,n}$ as the smallest integer such that for integers $x \geq R_{c,n}$, there are at least n primes between cx and x. Using consequences of strengthened versions of the Prime Number Theorem, we prove the existence of $R_{c,n}$ for all n and all c, that the asymptotic behavior is $R_{c,n} \sim p_{\frac{n}{1-c}}$ (where p_m is the mth prime), and that the percentage of primes that are c-Ramanujan converges to $1 - c$. We then study finer questions related to their distribution among the primes, and see that the $c-$Ramanujan primes display striking behavior, deviating significantly from a probabilistic model based on biased coin flipping. This model is related to the Cramer model, which correctly predicts many properties of primes on large scales but has been shown to fail in some instances on smaller scales. These results extend those of Sondow, Nicholson, and Noe, who proved and observed similar behavior for Ramanujan primes.

3.2 Results

Existence of $R_{c,n}$

For any $c \in (0, 1)$ and any positive integer n, the c-Ramanujan prime $R_{c,n}$ exists.

Asymptotic behavior of $R_{c,n}$

1. For any fixed $c \in (0, 1)$, the $n\mathrm{th}$ c-Ramanujan prime is asymptotic to the $\frac{n}{1-c}\mathrm{th}$ prime as $n \to \infty$, that is, $\lim_{n\to\infty} \frac{R_{c,n}}{p_{\frac{n}{1-c}}} = 1$. More precisely, there exists a constant $\beta_{1,c} > 0$ such that $|R_{c,n} - p_{\frac{n}{1-c}}| \leq \beta_{1,c}\, n \log\log n$ for all sufficiently large n.
2. In the limit, the probability of a generic prime being a c-Ramanujan prime is $1 - c$. More precisely, there exists a constant $\beta_{5,c}$ such that for N large we have $\left|\frac{\pi_c(N)}{\pi(N)} - (1 - c)\right| \leq \frac{\beta_{5,c} \log\log N}{\log N}$.

3.3 Distribution of generalized Ramanujan primes

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

	Length of the longest run in $(10^5, 10^6)$ of			
	c-Ramanujan primes		Non-Ramanujan primes	
c	Expected	Actual	Expected	Actual
0.05	127	97	4	2
0.10	70	58	5	3
0.15	49	42	6	6
0.20	38	36	7	7
0.25	30	27	9	12
0.30	25	25	10	12
0.35	21	18	11	18
0.40	18	21	13	16
0.45	16	19	14	23
0.50	14	20	16	36
0.55	12	16	19	39
0.60	11	17	22	42
0.65	10	13	25	53
0.70	9	14	30	78
0.75	8	11	37	119
0.80	7	9	46	154
0.85	6	10	62	303
0.90	5	11	91	345