Benford's law and Copulas

Thealexa Becker, Smith College

tbecker@smith.edu
http://www.williams.edu/Mathematics/sjmiller/

Smith College, September 24, 2011

Interesting Question

Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?

Interesting Question

Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?

Plausible answers:

Interesting Question

Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?

Plausible answers:10%

Interesting Question

Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?

Plausible answers:10%, 11%

Interesting Question

Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?

Plausible answers:10%, 11%, about 30%.

Summary

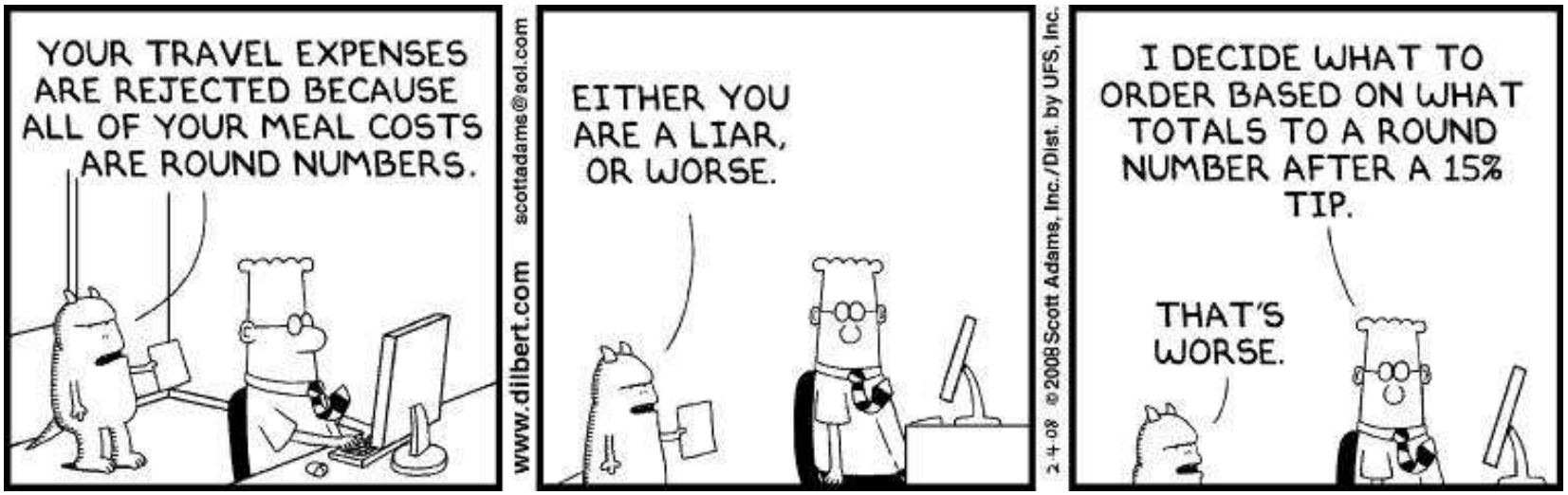
- State Benford's Law.
- Discuss examples and applications.
- Sketch proofs.
- Describe open problems.

Caveats!

- A math test indicating fraud is *not* proof of fraud: unlikely events, alternate reasons.

Caveats!

- A math test indicating fraud is *not* proof of fraud: unlikely events, alternate reasons.



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

Statement

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

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

Statement

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

- Not all data sets satisfy Benford's Law.

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

Statement

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

- Not all data sets satisfy Benford's Law.
  - ◇ Long street $[1, L]$: $L = 199$ versus $L = 999$.

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

Statement

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

- Not all data sets satisfy Benford's Law.
  - ◇ Long street $[1, L]$: $L = 199$ versus $L = 999$.
  - ◇ Oscillates between $1/9$ and $5/9$ with first digit 1.

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

Statement

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

- Not all data sets satisfy Benford's Law.
  - ◇ Long street $[1, L]$: $L = 199$ versus $L = 999$.
  - ◇ Oscillates between $1/9$ and $5/9$ with first digit 1.
  - ◇ Many streets of different sizes: close to Benford.

Examples

- recurrence relations
- special functions (such as $n!$)
- iterates of power, exponential, rational maps
- products of random variables
- L-functions, characteristic polynomials
- iterates of the $3x + 1$ map
- differences of order statistics
- hydrology and financial data
- many hierarchical Bayesian models

Applications

- analyzing round-off errors
- determining the optimal way to store numbers
- detecting tax and image fraud, and data integrity

General Theory

Mantissas

Mantissa: $x = M_{10}(x) \cdot 10^k$, k integer.

$M_{10}(x) = M_{10}(\widetilde{x})$ if and only if x and $\widetilde{x}$ have the same leading digits.

Key observation: $\log_{10}(x) = \log_{10}(\widetilde{x}) \bmod 1$ if and only if x and $\widetilde{x}$ have the same leading digits. Thus often study $y = \log_{10} x$.

Equidistribution and Benford's Law

Equidistribution

$\{y_n\}_{n=1}^{\infty}$ is equidistributed modulo 1 if probability $y_n \bmod 1 \in [a, b]$ tends to $b - a$:

$$\frac{\#\{n \leq N : y_n \bmod 1 \in [a, b]\}}{N} \rightarrow b - a.$$

Equidistribution and Benford's Law

Equidistribution

$\{y_n\}_{n=1}^{\infty}$ is equidistributed modulo 1 if probability $y_n \bmod 1 \in [a, b]$ tends to $b - a$:

$$\frac{\#\{n \leq N : y_n \bmod 1 \in [a, b]\}}{N} \rightarrow b - a.$$

- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.

Equidistribution and Benford's Law

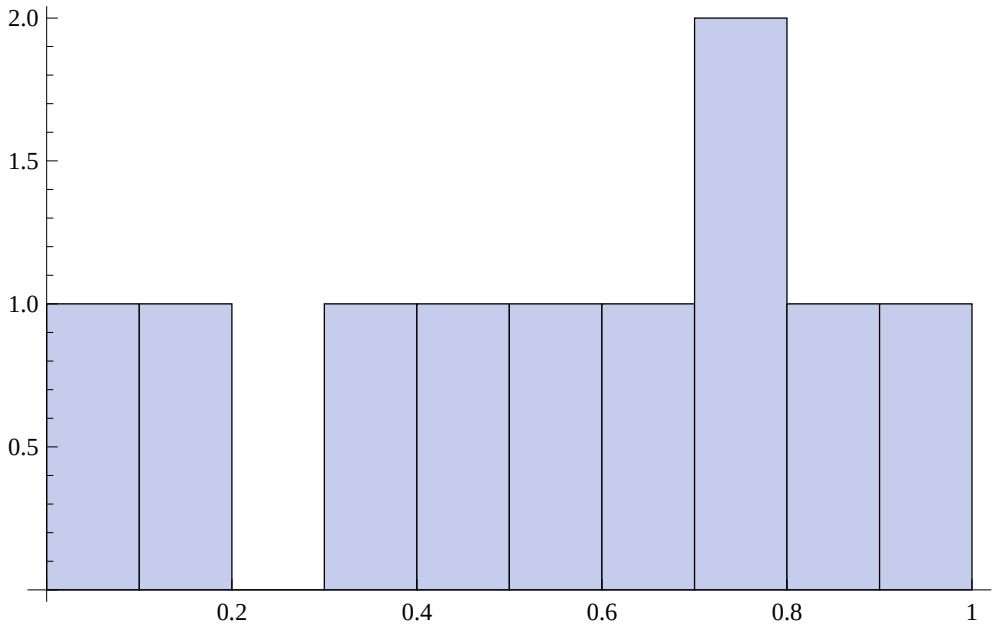
Equidistribution

$\{y_n\}_{n=1}^{\infty}$ is equidistributed modulo 1 if probability $y_n \bmod 1 \in [a, b]$ tends to $b - a$:

$$\frac{\#\{n \leq N : y_n \bmod 1 \in [a, b]\}}{N} \rightarrow b - a.$$

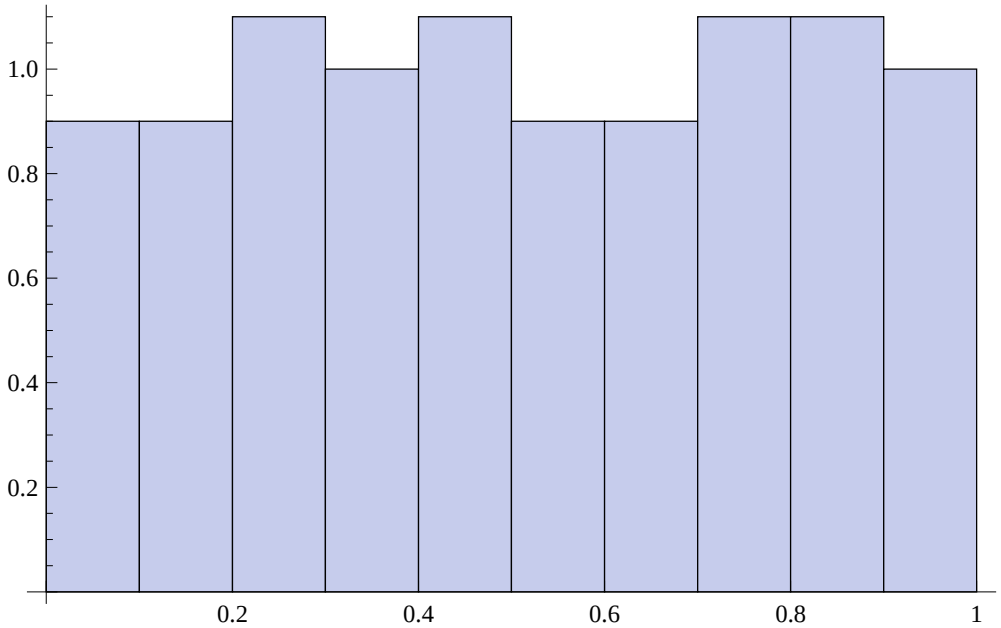
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.

Example of Equidistribution: $n\sqrt{\pi}$ mod 1



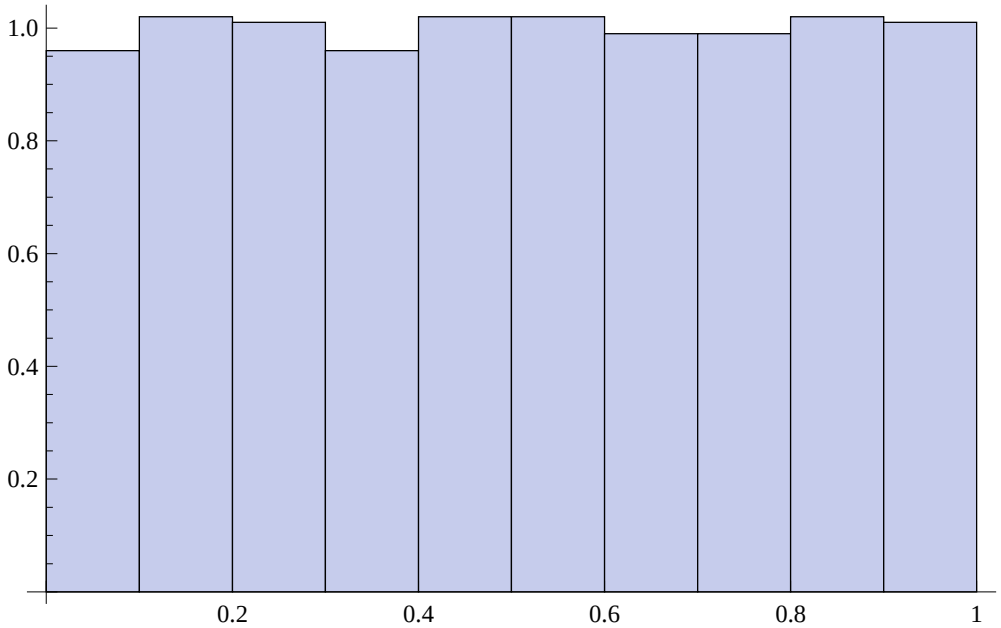
$n\sqrt{\pi}$ mod 1 for $n \leq 10$

Example of Equidistribution: $n\sqrt{\pi}$ mod 1



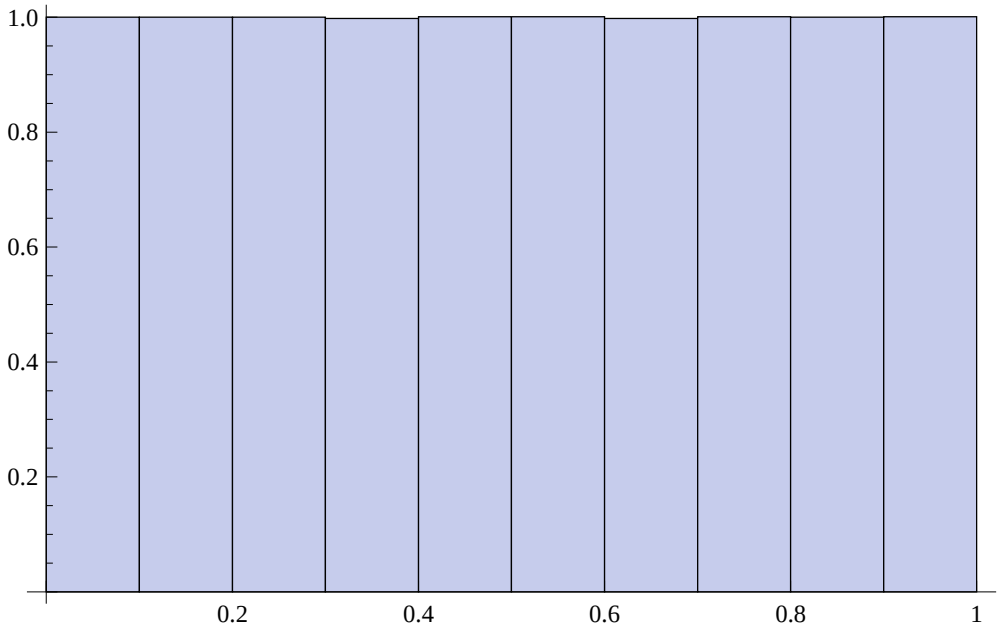
$n\sqrt{\pi}$ mod 1 for $n \leq 100$

Example of Equidistribution: $n\sqrt{\pi}$ mod 1



$n\sqrt{\pi}$ mod 1 for $n \leq 1000$

Example of Equidistribution: $n\sqrt{\pi}$ mod 1



$n\sqrt{\pi}$ mod 1 for $n \leq 10,000$

Logarithms and Benford's Law

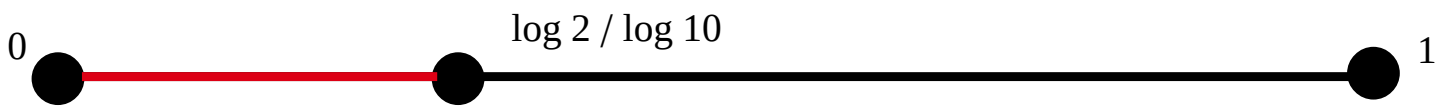
Fundamental Equivalence

Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.

Logarithms and Benford's Law

Fundamental Equivalence

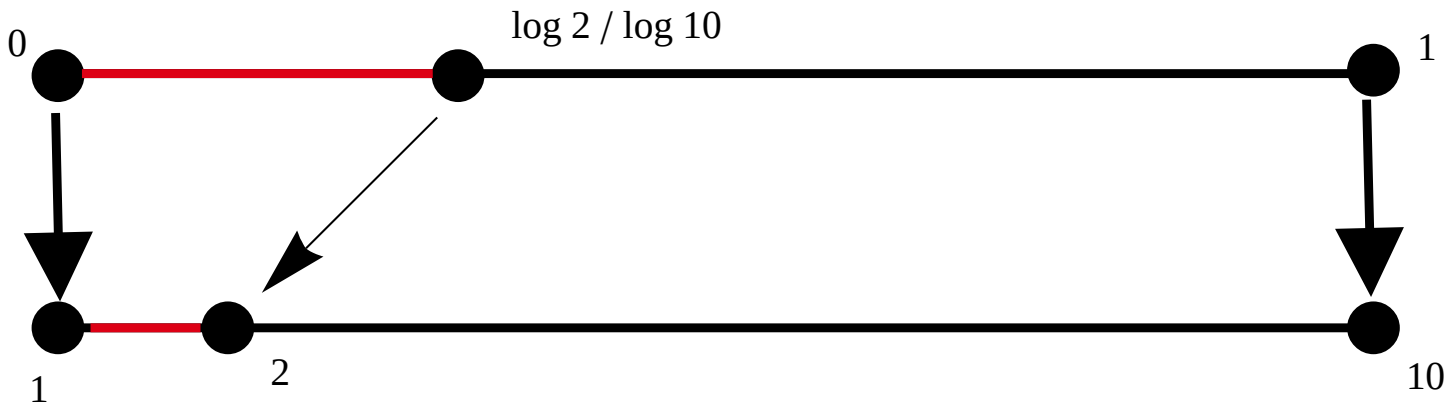
Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



Logarithms and Benford's Law

Fundamental Equivalence

Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



Examples

- 2^n is Benford base 10 as $\log_{10} 2 \notin \mathbb{Q}$.

Examples

- 2^n is Benford base 10 as $\log_{10} 2 \notin \mathbb{Q}$.

- Fibonacci numbers are Benford base 10.

  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = r^n$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n$.
  Approximation: $a_n \approx \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n$.

Detecting Fraud

Bank Fraud

- Audit of a bank revealed huge spike of numbers starting with

Detecting Fraud

Bank Fraud

- Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.

Detecting Fraud

Bank Fraud

- Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.
- Write-off limit of $5,000. Officer had friends applying for credit cards, ran up balances just under $5,000 then he would write the debts off.

Copulas and Benford's Law

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

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:

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

where

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

Conclusions

Conclusions and Future Investigations

- See many different systems exhibit Benford behavior.
- Ingredients of proofs (logarithms, equidistribution).
- Applications to fraud detection / data integrity.