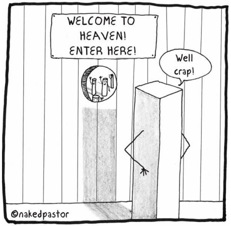
Benford's Law: Hammering a Square Peg Into a Round Hole?

Félix Balado and Guénolé C. Silvestre

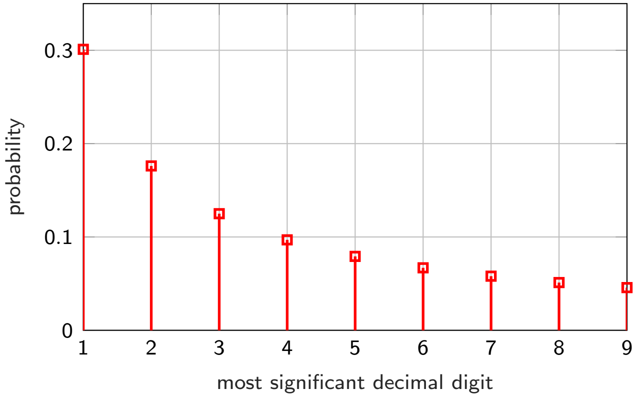
School of Computer Science
University College Dublin
Ireland

29th EUSIPCO, Dublin
23–27 August, 2021



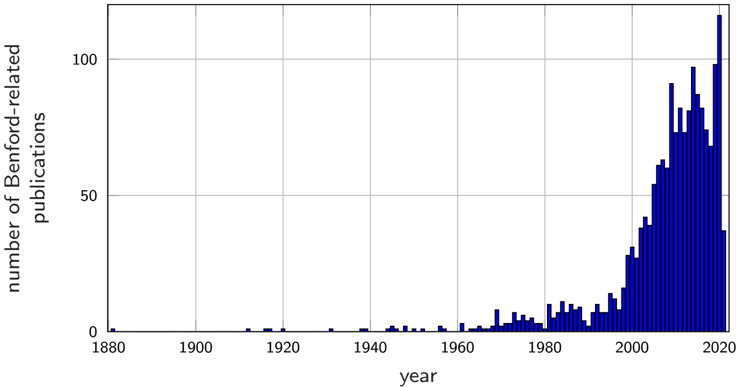
Benford's Law

- Newcomb (1881) and, independently, Benford (1938) noticed the following pattern in certain datasets:



Research on Benford's Law

- The appearance of Benford's distribution in many different scenarios has been extensively studied



total: 1,735 publications

[source: `benfordonline.net`]

Legal Disclaimer

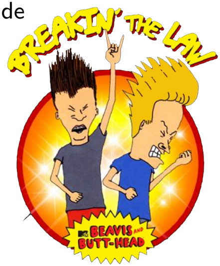
- Many recurrence relations comply exactly with Benford's law
  - Pochhammer numbers, Bell numbers, Fibonacci numbers...
  - reason: equidistribution theorem (Sierpiński, Weyl, c. 1909)

Legal Disclaimer

- Many recurrence relations comply exactly with Benford's law
  - Pochhammer numbers, Bell numbers, Fibonacci numbers...
  - reason: equidistribution theorem (Sierpiński, Weyl, c. 1909)
- But when it comes to data arising from natural random processes the justifications for Benford's law are shakier
  - e.g. Benford's law holds when
    - data exhibits geometric growth
    - data is spread over many orders of magnitude
    - *data is scale invariant*

Legal Disclaimer

- Many recurrence relations comply exactly with Benford's law
  - Pochhammer numbers, Bell numbers, Fibonacci numbers...
  - reason: equidistribution theorem (Sierpiński, Weyl, c. 1909)
- But when it comes to data arising from natural random processes the justifications for Benford's law are shakier
  - e.g. Benford's law holds when
    - data exhibits geometric growth
    - data is spread over many orders of magnitude
    - *data is scale invariant*



- Should we stop calling Benford's law a "law"?

General Distribution of the k Most Significant b-ary Digits

- Fractional part of $y \in \mathbb{R}$: $\{y\} = y - \lfloor y \rfloor$

General Distribution of the k Most Significant b-ary Digits

- Fractional part of $y \in \mathbb{R}$: $\{y\} = y - \lfloor y \rfloor$
- The discrete r.v. modelling the k most significant b-ary digits of a positive continuous r.v. X is

$$A_{(k)} = \lfloor b^{\{\log_b X\}+k-1} \rfloor, \text{ with support } \mathcal{A}_{(k)} = \{b^{k-1}, \dots, b^k - 1\}$$

General Distribution of the k Most Significant b-ary Digits

- Fractional part of $y \in \mathbb{R}$: $\{y\} = y - \lfloor y \rfloor$
- The discrete r.v. modelling the k most significant b-ary digits of a positive continuous r.v. X is

$$A_{(2)} = \lfloor 10^{\{\log_{10} X\}+2-1} \rfloor, \text{ with support } \mathcal{A}_{(2)} = \{10, 11, \ldots, 98, 99\}$$

General Distribution of the k Most Significant b-ary Digits

- Fractional part of $y \in \mathbb{R}$: $\{y\} = y - \lfloor y \rfloor$
- The discrete r.v. modelling the k most significant b-ary digits of a positive continuous r.v. X is

$$A_{(k)} = \lfloor b^{\{\log_b X\}+k-1} \rfloor, \text{ with support } \mathcal{A}_{(k)} = \{b^{k-1}, \ldots, b^k - 1\}$$

- Letting $Y = \log_b X$, the pmf of $A_{(k)}$ can be obtained from the cdf of $\{Y\}$, $F_{\{Y\}}(y) = \Pr(\{Y\} \leq y)$, as follows:

$$\boxed{\Pr(A_{(k)} = a) = F_{\{Y\}}\big(\log_b(a+1) - k + 1\big) - F_{\{Y\}}\big(\log_b a - k + 1\big)}$$

General Distribution of the k Most Significant b-ary Digits

- Fractional part of $y \in \mathbb{R}$: $\{y\} = y - \lfloor y \rfloor$
- The discrete r.v. modelling the k most significant b-ary digits of a positive continuous r.v. X is

$$A_{(k)} = \lfloor b^{\{\log_b X\}+k-1} \rfloor, \text{ with support } \mathcal{A}_{(k)} = \{b^{k-1}, \ldots, b^k - 1\}$$

- Letting $Y = \log_b X$, the pmf of $A_{(k)}$ can be obtained from the cdf of $\{Y\}$, $F_{\{Y\}}(y) = \Pr(\{Y\} \leq y)$, as follows:

$$\boxed{\Pr(A_{(k)} = a) = F_{\{Y\}}\big(\log_b(a+1) - k + 1\big) - F_{\{Y\}}\big(\log_b a - k + 1\big)}$$

  - the j-th MSD can also be modelled using $A_{[j]} = A_{(j)} \pmod{b}$

Getting Particular

- Definition: X is Benford if $\{Y\} \sim U(0,1) \Rightarrow F_{\{Y\}}(y) = y$

Getting Particular

- Definition: X is Benford if $\{Y\} \sim U(0,1) \Rightarrow F_{\{Y\}}(y) = y$
  - general expression leads to well-known Benford's distribution

$$\Pr(A_{(k)} = a) = \log_b \left(1 + \frac{1}{a}\right), \quad \text{where } a \in \mathcal{A}_{(k)}$$

Getting Particular

- Definition: X is Benford if $\{Y\} \sim U(0,1) \Rightarrow F_{\{Y\}}(y) = y$
  - general expression leads to well-known Benford's distribution

$$\Pr(A_{(2)} = 45) = \log_{10}\left(1 + \frac{1}{45}\right)$$

Getting Particular

- Definition: X is Benford if $\{Y\} \sim U(0,1) \Rightarrow F_{\{Y\}}(y) = y$
  - general expression leads to well-known Benford's distribution

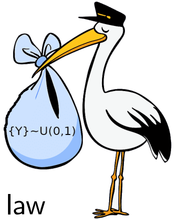
$$\Pr(A_{(k)} = a) = \log_b\left(1 + \frac{1}{a}\right), \quad \text{where } a \in \mathcal{A}_{(k)}$$

  - the j-th MSD (for $j \geq 2$) is distributed as

$$\Pr(A_{[j]} = a) = \log_b\left(\frac{\Gamma\left((a+1)b^{-1} + b^{j-1}\right)\Gamma\left(ab^{-1} + b^{j-2}\right)}{\Gamma\left((a+1)b^{-1} + b^{j-2}\right)\Gamma\left(ab^{-1} + b^{j-1}\right)}\right)$$

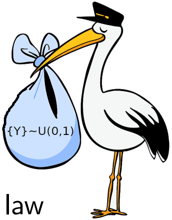
  where $a \in \{0, 1, \ldots, b-1\}$ and $\Gamma(\cdot)$ is the Gamma function
    - this closed-form expression was never previously given

But. . . Where Do Benford r.v.'s Come From?

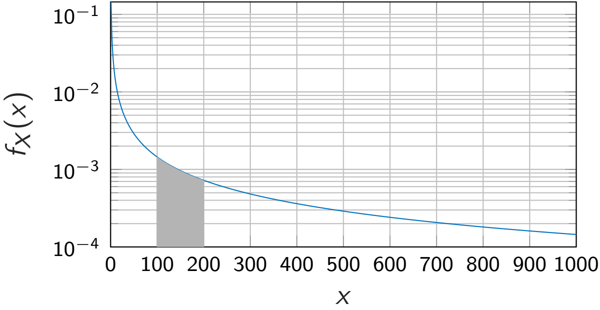


- Pinkham (1961): scale invariance is behind Benford's law

But. . . Where Do Benford r.v.'s Come From?

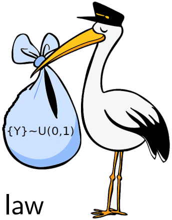


- Pinkham (1961): scale invariance is behind Benford's law

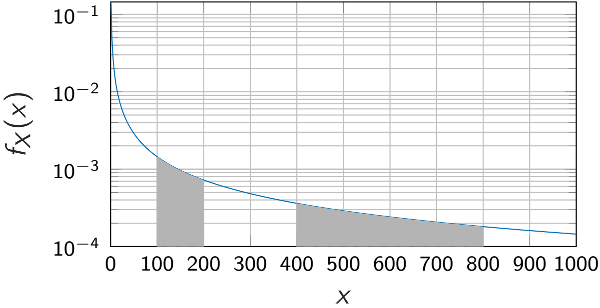


$\Pr\left(X \in (100, 200)\right)$

But...Where Do Benford r.v.'s Come From?

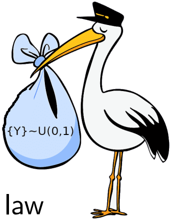


- Pinkham (1961): scale invariance is behind Benford's law

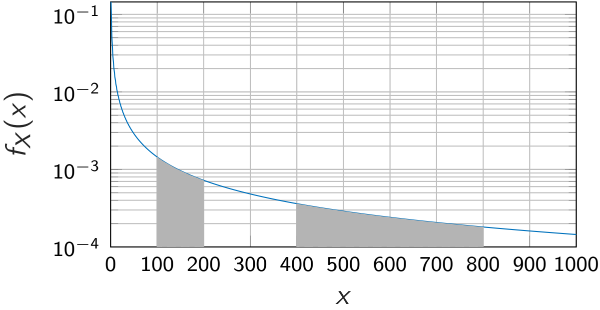


$$\Pr\left(X \in (100, 200)\right) = \Pr\left(X \in 4 \times (100, 200)\right)$$

But. . . Where Do Benford r.v.'s Come From?



- Pinkham (1961): scale invariance is behind Benford's law



$$\Pr\left(X \in (x', x)\right) = \Pr\left(X \in \alpha\,(x', x)\right) \Rightarrow X \text{ is strictly scale invariant}$$

Strict Scale Invariance and Base Invariance

- Property of the pdf of strictly scale-invariant X

$$\boxed{f_X(x) = \alpha\, f_X(\alpha\, x)} \quad \alpha > 0$$

Strict Scale Invariance and Base Invariance

- Property of the pdf of strictly scale-invariant X

$$\boxed{f_X(x) = \alpha\, f_X(\alpha\, x)} \quad \alpha > 0$$

- Consequences: Y is uniform, and so X must have finite support which must also depend on b to ensure $\{Y\} \sim U(0,1)$

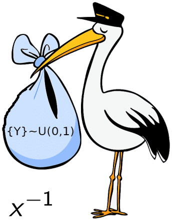
→ The common notion "*scale-invariant data that follows Benford's law is base invariant*" can only be an approximation

The One and Only, But Often a Misfit

- The pdf of a strictly scale invariant r.v. X must be $\propto x^{-1}$
  $\rightarrow$ the prize-competition distribution is the only choice

$$f_X(x) = \frac{1}{x \ln(x_\mathrm{M}/x_\mathrm{m})}, \quad 0 < x_\mathrm{m} \leq x \leq x_\mathrm{M}$$

The One and Only, But Often a Misfit

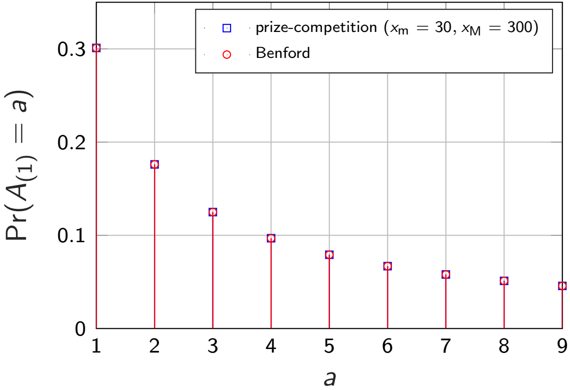


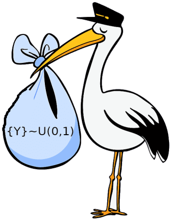
- The pdf of a strictly scale invariant r.v. X must be $\propto x^{-1}$
  $\rightarrow$ the prize-competition distribution is the only choice

$$f_X(x) = \frac{1}{x \ln(x_\mathrm{M}/x_\mathrm{m})}, \quad 0 < x_\mathrm{m} \leq x \leq x_\mathrm{M}$$

  - plus, for X to be Benford it must hold that $\log_b(x_\mathrm{M}/x_\mathrm{m}) \in \mathbb{Z}$

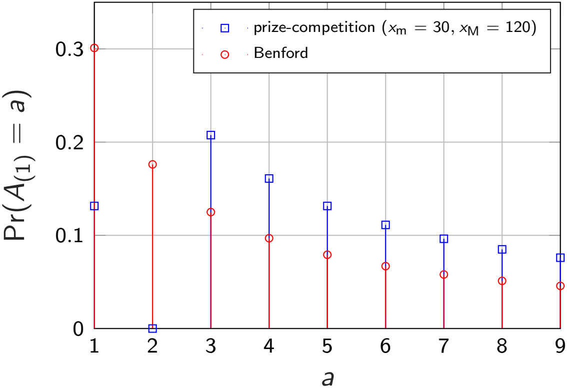
First Significant Digit in Prize-Competition Distribution





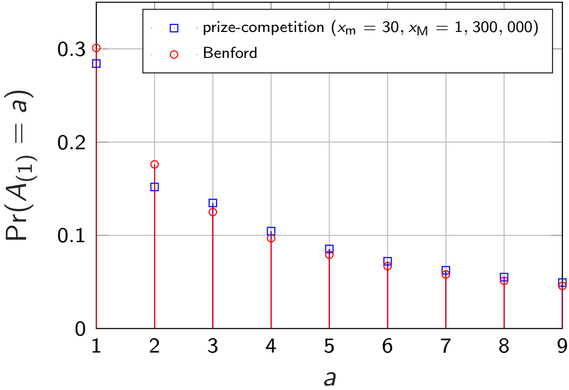
- If $\log_b(x_M/x_m) \in \mathbb{Z}$ we get Benford's distribution

First Significant Digit in Prize-Competition Distribution



- If $\log_b(x_M/x_m) \notin \mathbb{Z}$ a mismatch is inevitable...

First Significant Digit in Prize-Competition Distribution



- If $\log_b(x_M/x_m) \notin \mathbb{Z}$ a mismatch is inevitable. . . but it decreases if the pdf spreads over many orders of magnitude
- Still, the prize-competition distribution is relatively uncommon

More Plausible Scale Invariance

- Consider a more relaxed definition of scale invariance:

$$\boxed{f_X(x) = \alpha^{\nu} f_X(\alpha\, x)} \quad \nu > 1$$

$\rightarrow$ The Pareto pdf is the only one to conform to this criterion

$$f_X(x) = \frac{s\, x_{\mathrm{m}}^s}{x^{s+1}}, \quad 0 < x_{\mathrm{m}} \le x,\ s > 0$$

More Plausible Scale Invariance

- Consider a more relaxed definition of scale invariance:

$$\boxed{f_X(x) = \alpha^{\nu} f_X(\alpha\, x)} \quad \nu > 1$$

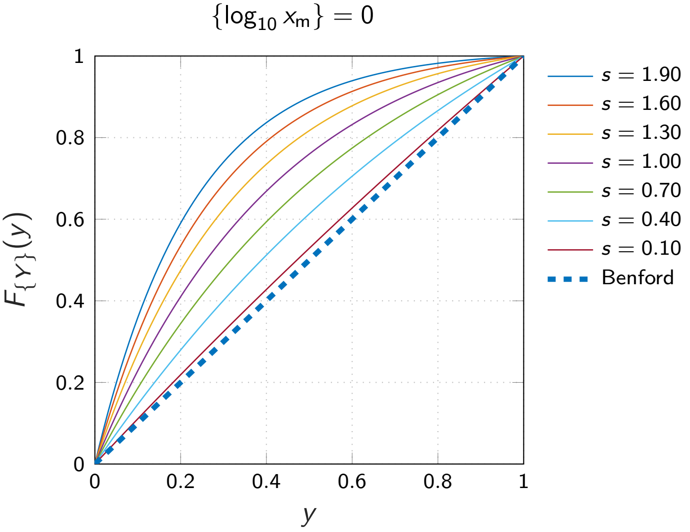
$\rightarrow$ The Pareto pdf is the only one to conform to this criterion

$$f_X(x) = \frac{s\, x_{\mathrm{m}}^{s}}{x^{s+1}}, \quad 0 < x_{\mathrm{m}} \leq x,\ s > 0$$

- Relevance: the Central Limit Theorem has a hidden side...
  - "heavy-tailed distributions, such as Pareto, are as prominent as the Gaussian distribution —if not more" (Nair et al., 2021)

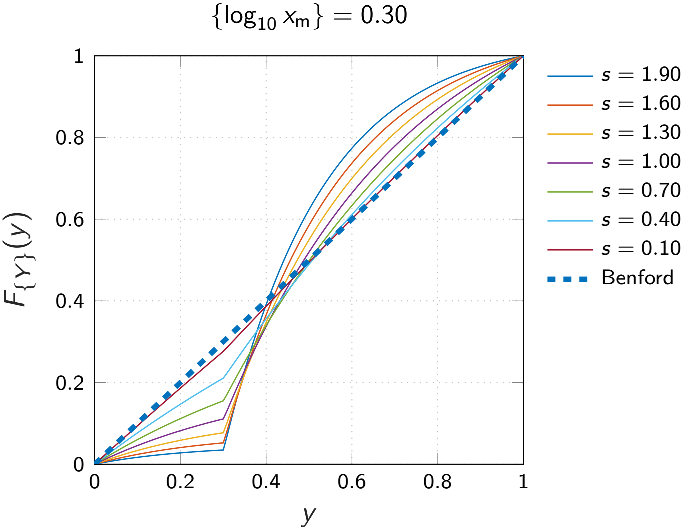
cdf of $\{Y\} = \{\log_b X\}$ for Pareto X

s: shape parameter
x_m: minimum value



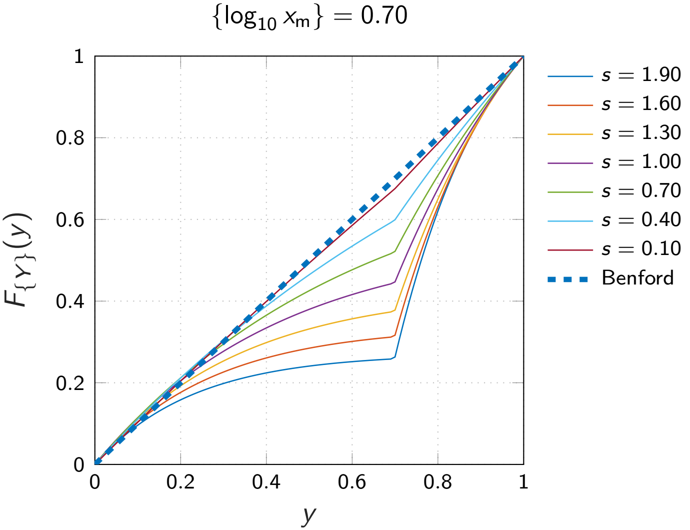
cdf of $\{Y\} = \{\log_b X\}$ for Pareto X

s: shape parameter
x_m: minimum value



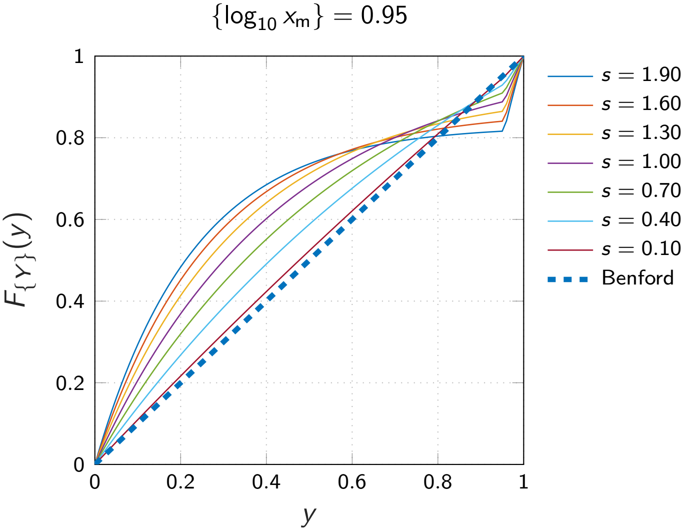
cdf of $\{Y\} = \{\log_b X\}$ for Pareto X

s: shape parameter
x_{m}: minimum value



cdf of $\{Y\} = \{\log_b X\}$ for Pareto X

s: shape parameter
x_{m}: minimum value



Wrapping it Up

- With the cdf of $\{Y\}$ and the general expression, we get the pmf of the k most significant b-ary digits for a Pareto r.v. X

$$\boxed{\begin{aligned}\Pr(A_{(k)} = a) = \frac{b^{s(\xi-1)}}{1-b^{-s}} & \left(a^{-s} - (a+1)^{-s}\right) \\ & + u\big(a+1-b^{\xi}\big)\big(1-b^{s\,\xi}(a+1)^{-s}\big) \\ & - u\big(a-b^{\xi}\big)\big(1-b^{s\,\xi}\,a^{-s}\big)\end{aligned}}$$

where $a \in \mathcal{A}_{(k)}$, $\xi = \{\log_b x_{\mathrm{m}}\} + k - 1$ and $u(\cdot)$ is unit-step function

Wrapping it Up

- With the cdf of $\{Y\}$ and the general expression, we get the pmf of the k most significant b-ary digits for a Pareto r.v. X

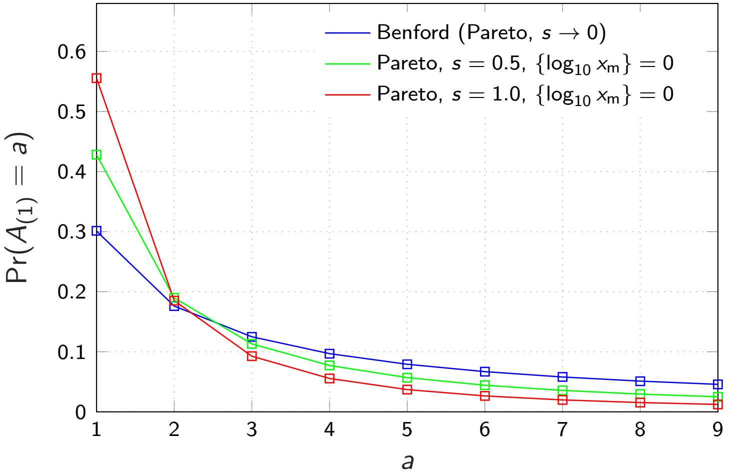
$$\boxed{\begin{aligned}\Pr(A_{(k)} = a) = \frac{b^{s(\xi-1)}}{1-b^{-s}}\left(a^{-s} - (a+1)^{-s}\right) \\ + u\left(a+1-b^{\xi}\right)\left(1 - b^{s\,\xi}(a+1)^{-s}\right) \\ - u\left(a-b^{\xi}\right)\left(1 - b^{s\,\xi}\,a^{-s}\right)\end{aligned}}$$

where $a \in \mathcal{A}_{(k)}$, $\xi = \{\log_b x_{\mathrm{m}}\} + k - 1$ and $u(\cdot)$ is unit-step function

  - as $s \to 0$ the distribution above tends to Benford's
  - **but**: the significant digits of scale-invariant datasets are far more likely to follow this distribution rather than Benford's

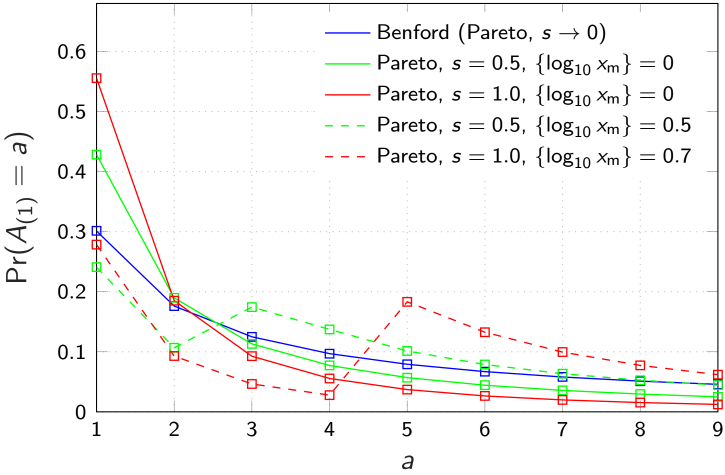
Distribution of the k MSDs of a Pareto Variable

- Pseudorandom empiricals vs theoreticals, $k = 1$



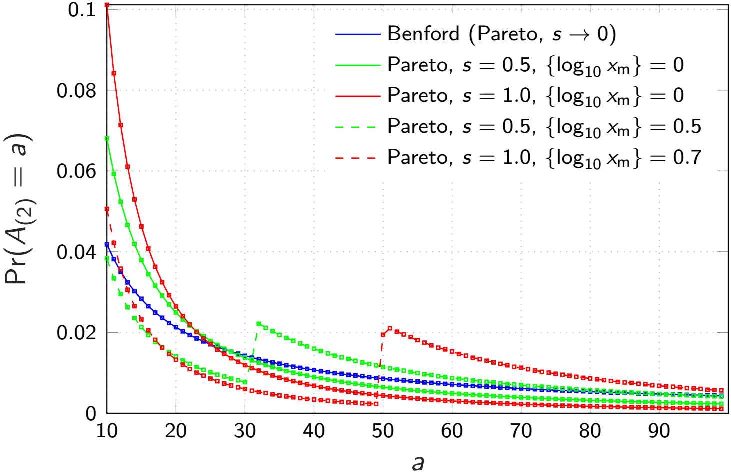
Distribution of the k MSDs of a Pareto Variable

- Pseudorandom empiricals vs theoreticals, $k = 1$



Distribution of the k MSDs of a Pareto Variable

- Pseudorandom empiricals vs theoreticals, $k = 2$



The Butterfly Effect

- Special case $\{\log_b x_\mathrm{m}\} = 0$ (i.e. no kink in the pmf)

$$\Pr(A_{(k)} = a) = \frac{a^{-s} - (a+1)^{-s}}{b^{-s(k-1)} - b^{-sk}}, \qquad a \in \mathcal{A}_{(k)}$$

  - originally found by Pietronero et al. (2001) for $k = 1$, then extended to general k by Barabesi and Pratelli (2020)

The Butterfly Effect

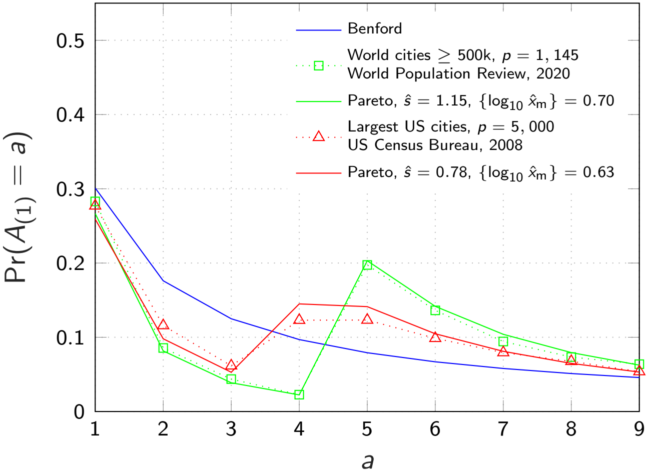
- Special case $\{\log_b x_\mathrm{m}\} = 0$ (i.e. no kink in the pmf)

$$\Pr(A_{(k)} = a) = \frac{a^{-s} - (a+1)^{-s}}{b^{-s(k-1)} - b^{-sk}}, \qquad a \in \mathcal{A}_{(k)}$$

  - originally found by Pietronero et al. (2001) for $k = 1$, then extended to general k by Barabesi and Pratelli (2020)
- Identified and named only in 2015, in a Lepidoptera study by Kozubowski et al.: discrete truncated Pareto (DTP) pmf
  - jaw-dropping fact: DTP can be obtained by quantising either
    1. a truncated Pareto r.v.
    2. the fractional part of the logarithm of a standard Pareto r.v.

First Significant Digit in Real Scale-Invariant Datasets

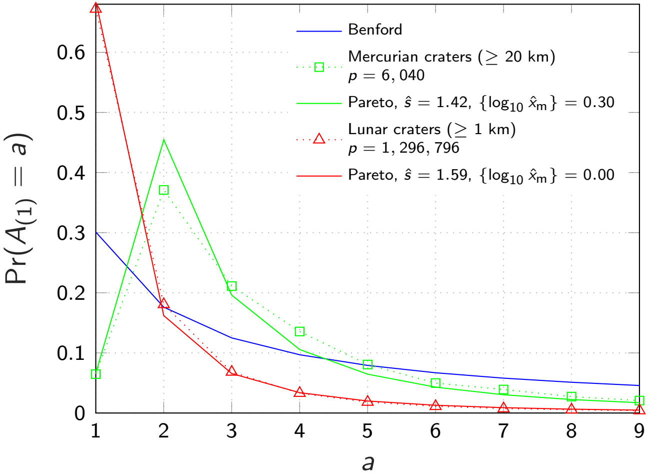
- Scale-invariant datasets are typically assumed to follow Benford's distribution...

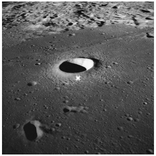


$\hat{s}$, $\hat{x}_\mathrm{m}$: ML estimators; p: dataset size

First Significant Digit in Real Scale-Invariant Datasets

- Scale-invariant datasets are typically assumed to follow Benford's distribution. . .

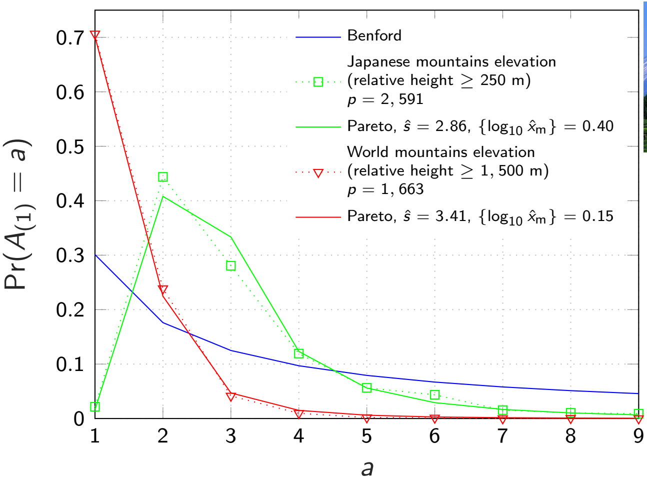




$\hat{s}$, $\hat{x}_m$: ML estimators; p: dataset size

First Significant Digit in Real Scale-Invariant Datasets

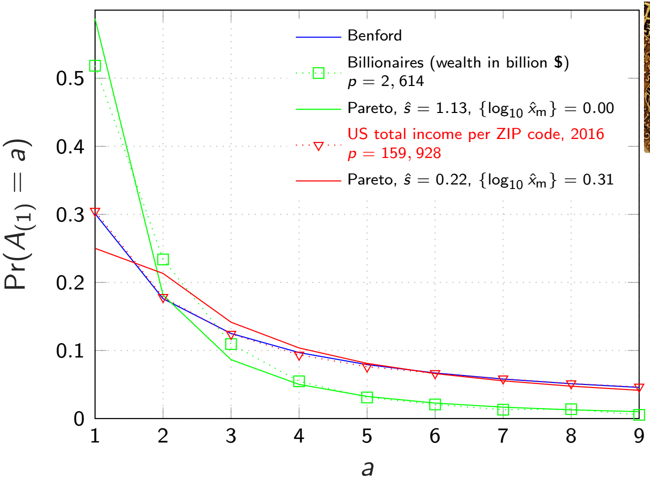
- Scale-invariant datasets are typically assumed to follow Benford's distribution...



$\hat{s}$, $\hat{x}_\mathrm{m}$: ML estimators; p: dataset size

First Significant Digit in Real Scale-Invariant Datasets

- Scale-invariant datasets are typically assumed to follow Benford's distribution. . . and sometimes they do!



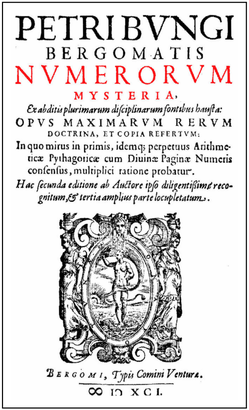
$\hat{s}$, $\hat{x}_m$: ML estimators; p: dataset size

What is the Significance of Significant Digits?

- The quintessential application of MSDs modelling is forensic analysis
  - tampering detection in economic data, election results, multimedia, etc

What is the Significance of Significant Digits?

- The quintessential application of MSDs modelling is forensic analysis
    - tampering detection in economic data, election results, multimedia, etc
- But: why look at the most significant digits of a set of numbers instead of looking at those numbers themselves?


PETRI BVNGI
BERGOMATIS
NVMERORVM
MYSTERIA,
Ex abditis plurimarum disciplinarum fontibus hausta:
OPVS MAXIMARVM RERVM
DOCTRINA, ET COPIA REFERTVM:
In quo mirus in primis, idemq; perpetuus Arithmeticæ Pythagoricæ cum Diuinæ Paginæ Numeris consensus, multiplici ratione probatur.
Hac secunda editione ab Auctore ipso diligentissime recognitum, & tertia amplius parte locupletatum.
BERGOMI, *Typis Comini Venturæ.*
∞ IↃ XCI.

Chasing Shadows in Forensic Analysis...

- Discrete projection of continuous data → <u>information loss</u>

Plato's allegory of the cave

"light came into the world, and men loved darkness rather than light"

Chasing Shadows in Forensic Analysis. . .

- Discrete projection of continuous data → information loss

Plato's allegory of the cave

"light came into the world, and men loved darkness rather than light"

Chasing Shadows in Forensic Analysis...

- Discrete projection of continuous data → information loss

Plato's allegory of the cave

"light came into the world, and men loved darkness rather than light"

Your Significant Others: Continued Fraction Coefficients

- Continued fractions (CF): a way of representing numbers alternative to positional base b number systems

$$y_0 = \lfloor y_0 \rfloor + \{y_0\}$$

Your Significant Others: Continued Fraction Coefficients

- Continued fractions (CF): a way of representing numbers alternative to positional base b number systems

$$y_0 = a_0 + \frac{1}{y_1}$$

Your Significant Others: Continued Fraction Coefficients

- Continued fractions (CF): a way of representing numbers alternative to positional base b number systems

$$y_0 = a_0 + \frac{1}{\lfloor y_1 \rfloor + \{y_1\}}$$

Your Significant Others: Continued Fraction Coefficients

- Continued fractions (CF): a way of representing numbers alternative to positional base b number systems

$$y_0 = a_0 + \frac{1}{a_1 + \frac{1}{y_2}}$$

Your Significant Others: Continued Fraction Coefficients

- Continued fractions (CF): a way of representing numbers alternative to positional base b number systems

$$y_0 = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}}$$

Your Significant Others: Continued Fraction Coefficients

- Continued fractions (CF): a way of representing numbers alternative to positional base b number systems

$$y_0 = [a_0; a_1, a_2, a_3, \ldots]$$

Your Significant Others: Continued Fraction Coefficients

- Continued fractions (CF): a way of representing numbers alternative to positional base b number systems

$$Y_0 = [A_0; A_1, A_2, A_3, \ldots]$$

Your Significant Others: Continued Fraction Coefficients

- Continued fractions (CF): a way of representing numbers alternative to positional base b number systems

$$Y_0 = [A_0; A_1, A_2, A_3, \ldots]$$

- If $Y_0 = \log_b X$ and X is Benford, then

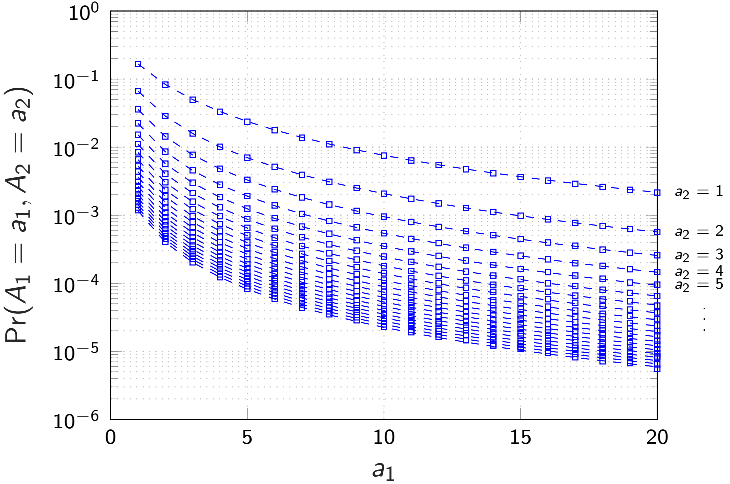
$$\boxed{\begin{aligned}\Pr(A_1 = a_1, \ldots, A_k = a_k) =& (-1)^k \big([0; a_1, \ldots, a_{k-1}, a_k + 1] \\ & - [0; a_1, \ldots, a_{k-1}, a_k]\big)\end{aligned}}$$

where $a_j \in \mathbb{N}$

$\rightarrow$ model for k most significant CF coefficients of $\log_b X$, analogous to model for k most significant b-ary digits of X

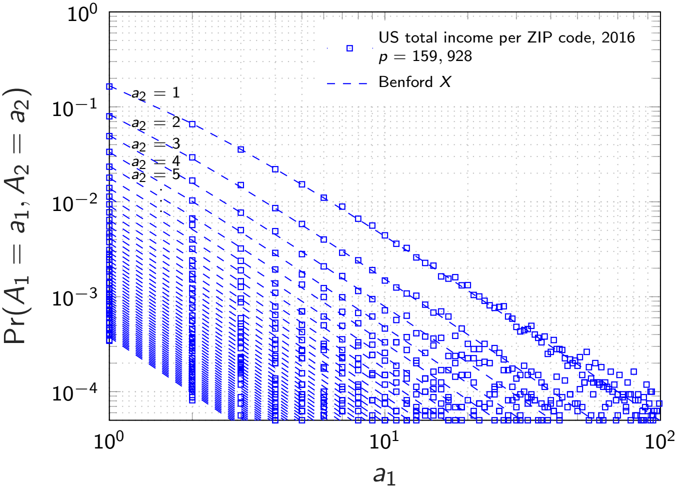
Distribution of the Two Most Significant CF Coefficients

- Pseudorandom empiricals vs theoreticals (Benford X)



CF Coefficients in Real Scale-Invariant Datasets

- Distribution of first two CF coefficients of $\log_{10} x_i$



p: dataset size

First CF Coefficient A_1 vs First Significant b-ary Digit $A_{(1)}$

A_1  $A_{(1)}$

- Which r.v. should we use in a forensic detection test where X is hypothesised to be Benford?
  a) $A_1 \;\; = \lfloor \{\log_b X\}^{-1} \rfloor$
  b) $A_{(1)} = \lfloor b^{\{\log_b X\}} \rfloor$

First CF Coefficient A_1 vs First Significant b-ary Digit $A_{(1)}$

$\boxed{A_1}$  $\cancel{A_{(1)}}$

- Which r.v. should we use in a forensic detection test where X is hypothesised to be Benford?
  - a) $A_1 \;\; = \lfloor \{\log_b X\}^{-1} \rfloor$
  - b) $A_{(1)} = \lfloor b^{\{\log_b X\}} \rfloor$
- Possible answers:
  - a) because there is <u>less information loss</u> wrt $\{Y\} = \{\log_b X\}$

$$\begin{aligned} I(A_1;\{Y\}) &= 2.046 \text{ nats} \\ I(A_{(1)};\{Y\}) &= 1.993 \text{ nats} \quad (b = 10) \end{aligned}$$

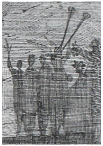
First CF Coefficient A_1 vs First Significant b-ary Digit $A_{(1)}$

$\cancel{A_1}$  $\boxed{A_{(1)}}$

- Which r.v. should we use in a forensic detection test where X is hypothesised to be Benford?
  - a) $A_1 \; = \lfloor \{\log_b X\}^{-1} \rfloor$
  - b) $A_{(1)} = \lfloor b^{\{\log_b X\}} \rfloor$
- Possible answers:
  - b) because there is less information loss wrt $\{Y\} = \{\log_b X\}$

$$
\begin{aligned}
I(A_1; \{Y\}) &= 2.046 \text{ nats} \\
I(A_{(1)}; \{Y\}) &= 2.413 \text{ nats} \quad (b = 16)
\end{aligned}
$$

First CF Coefficient A_1 vs First Significant b-ary Digit $A_{(1)}$

$\cancel{A_1}$  $\cancel{A_{(1)}}$

- Which r.v. should we use in a forensic detection test where X is hypothesised to be Benford?
  - a) $A_1 \;\; = \lfloor \{\log_b X\}^{-1} \rfloor$
  - b) $A_{(1)} = \lfloor b^{\{\log_b X\}} \rfloor$
- Possible answers:
  - $\rightarrow$ none of them: using $\{\log_b X\}$ should always be better

Time to Recap

1. The most significant digits in scale-invariant data can often be modelled using a generalisation of Benford's distribution based on heavy-tailed Pareto variables
2. There is nothing special about significant b-ary digits: they may be replaced by significant continued fraction coefficients in forensic detection tests

Time to Recap

1. The most significant digits in scale-invariant data can often be modelled using a generalisation of Benford's distribution based on heavy-tailed Pareto variables
2. There is nothing special about significant *b*-ary digits: they may be replaced by significant continued fraction coefficients in forensic detection tests
   - and both are just shadows. . .

Go raibh míle maith agaibh