Zipf's law and friends

Hera He

January 8, 2016

Outline

1. Power law
2. Zipf's law
3. Heap's law
4. Benford's law

Normal distribution

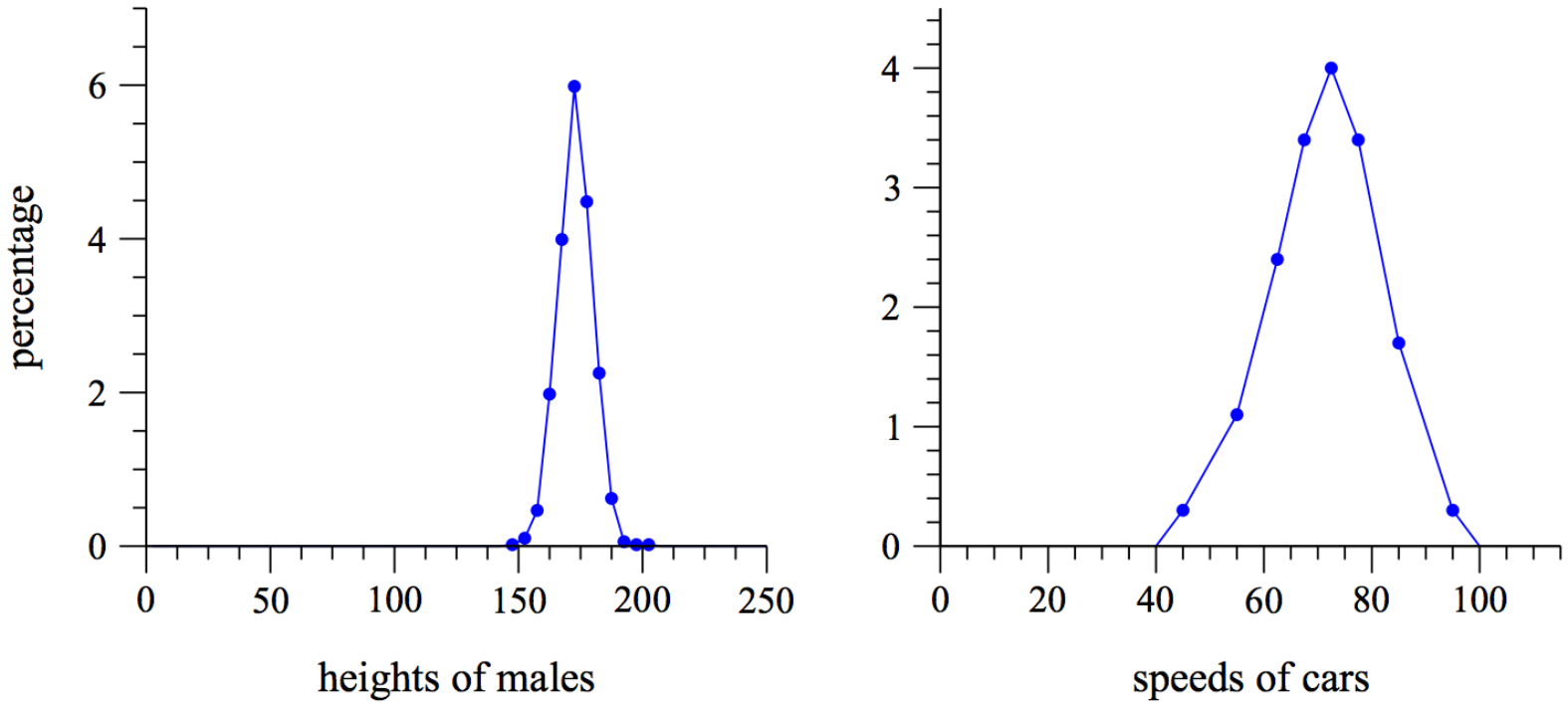


Figure : Left: histogram of heights in centimetres of American males. Data from the National Health Examination Survey, 19591962 (US Department of Health and Human Services). Right: histogram of speeds in miles per hour of cars on UK motorways. Data from Transport Statistics 2003 (UK Department for Transport).

Power law: first example

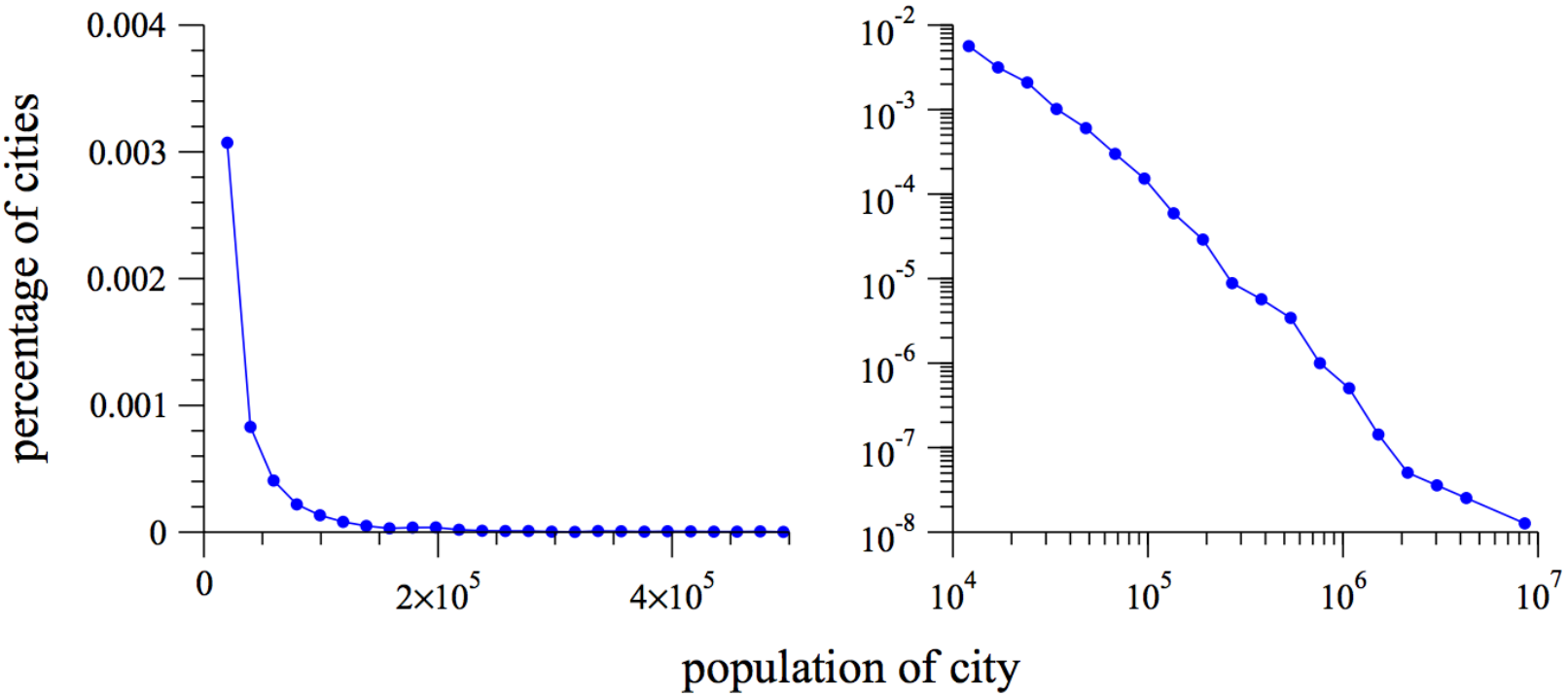


Figure : Left: histogram of the populations of all US cities with population of 10,000 or more. Right: another histogram of the same data, but plotted on logarithmic scales. The approximate straight-line form of the histogram in the right panel implies that the distribution follows a power law. Data from the 2000 US Census.

Power law: Definition

$$\ln p(x) = -\alpha \ln x + c$$
$$\Leftrightarrow p(x) = Cx^{-\alpha}$$

- Distributions of the above form are said to follow a power law.
- The constant α is called the **exponent** of the power law.
- Other names: Patero distribution(continuous), Zipf's law(discrete).

Measuring power law

How to detect or identify power-law behaviour?

First thought: make a simple histogram and plot it on log scale.

Poor way to proceed!

Measuring power law

We simulate 10^6 random numbers from a power-law probability with exponent $\alpha = -2.5$.(plot a) Simple histogram plotted on log-scale produces noisy fluctuation in the tail.(plot b)

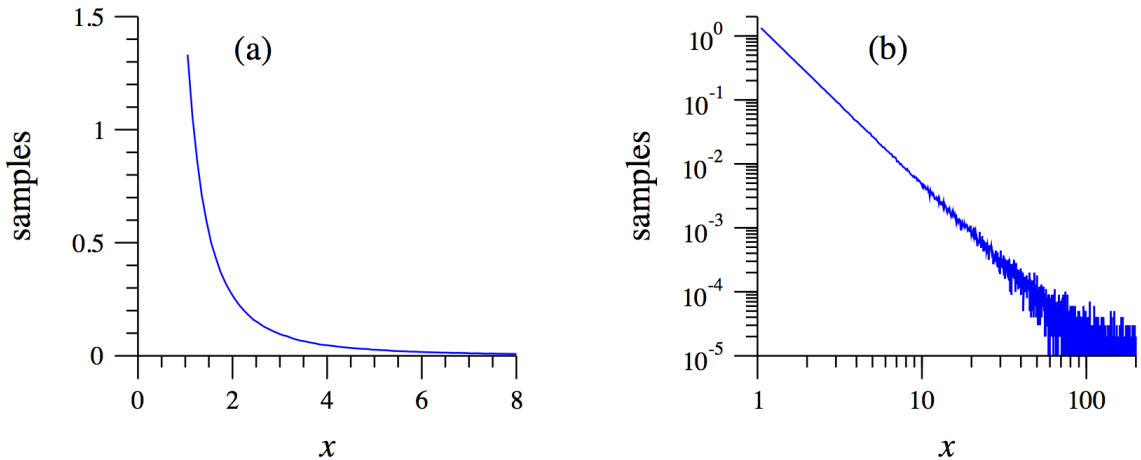


Figure : (a) Histogram of the set of 1 million random numbers described in the text, which have a power-law distribution with exponent $\alpha = 2.5$. (b) The same histogram on logarithmic scales. Notice how noisy the results get in the tail towards the right-hand side of the panel. This happens because the number of samples in the bins becomes small and statistical fluctuations are therefore large as a fraction of sample number.

Measuring power law

A better plot with logrightmic binning: each bin is a fixed multiplier wider than the one before. e.g. use bins of size 0.1, 0.2, 0.4, 0.8 etc. (break points (1, 1.1), (1.1, 1.3), (1.3, 1.7), (1.7, 2.5) etc.)

Measuring power law

A better plot with logrightmic binning: each bin is a fixed multiplier wider than the one before. e.g. use bins of size 0.1, 0.2, 0.4, 0.8 etc. (break points (1, 1.1), (1.1, 1.3), (1.3, 1.7), (1.7, 2.5) etc.)

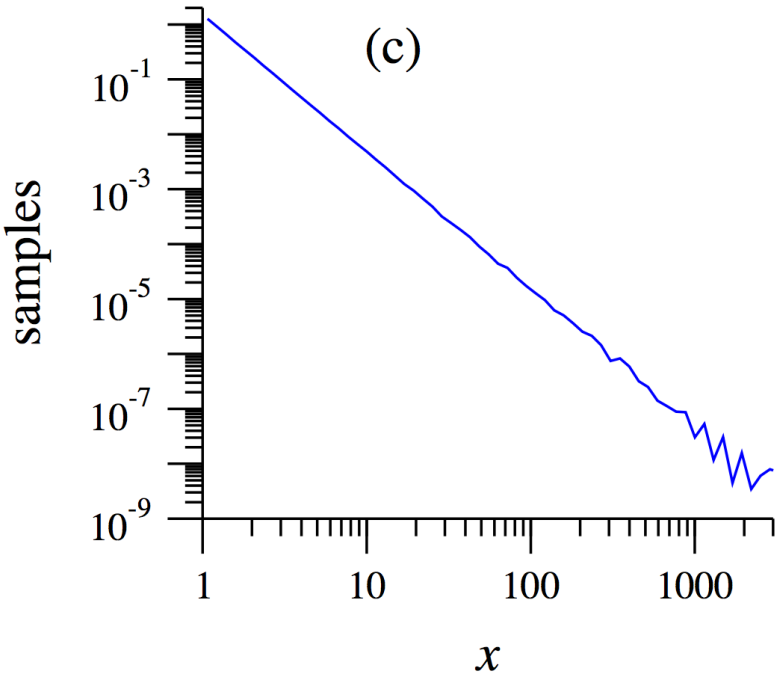


Figure : (c) A histogram constructed using logarithmic binning.

Measuring power law

A superior plotting method: Plot ccdf(tail probability) $P(X > x)$ against data values.
Or equivalently, use rank/frequency plot.

$$P(X > x) = \int_x^\infty p(t)dt = \frac{C}{\alpha - 1} x^{-(\alpha-1)}$$

Measuring power law

A superior plotting method: Plot ccdf(tail probability) $P(X > x)$ against data values.
Or equivalently, use rank/frequency plot.

$$P(X > x) = \int_x^\infty p(t)dt = \frac{C}{\alpha - 1} x^{-(\alpha-1)}$$

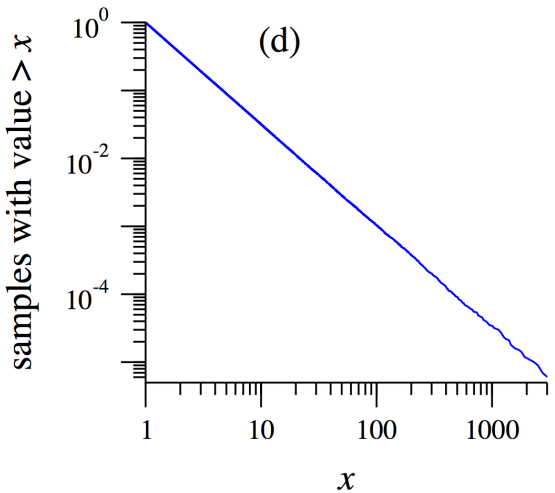


Figure : (d) A cumulative histogram or rank/frequency plot of the same data. The cumulative distribution also follows a power law, but with an exponent of $\alpha - 1 = 1.5$.

Estimation of the exponent

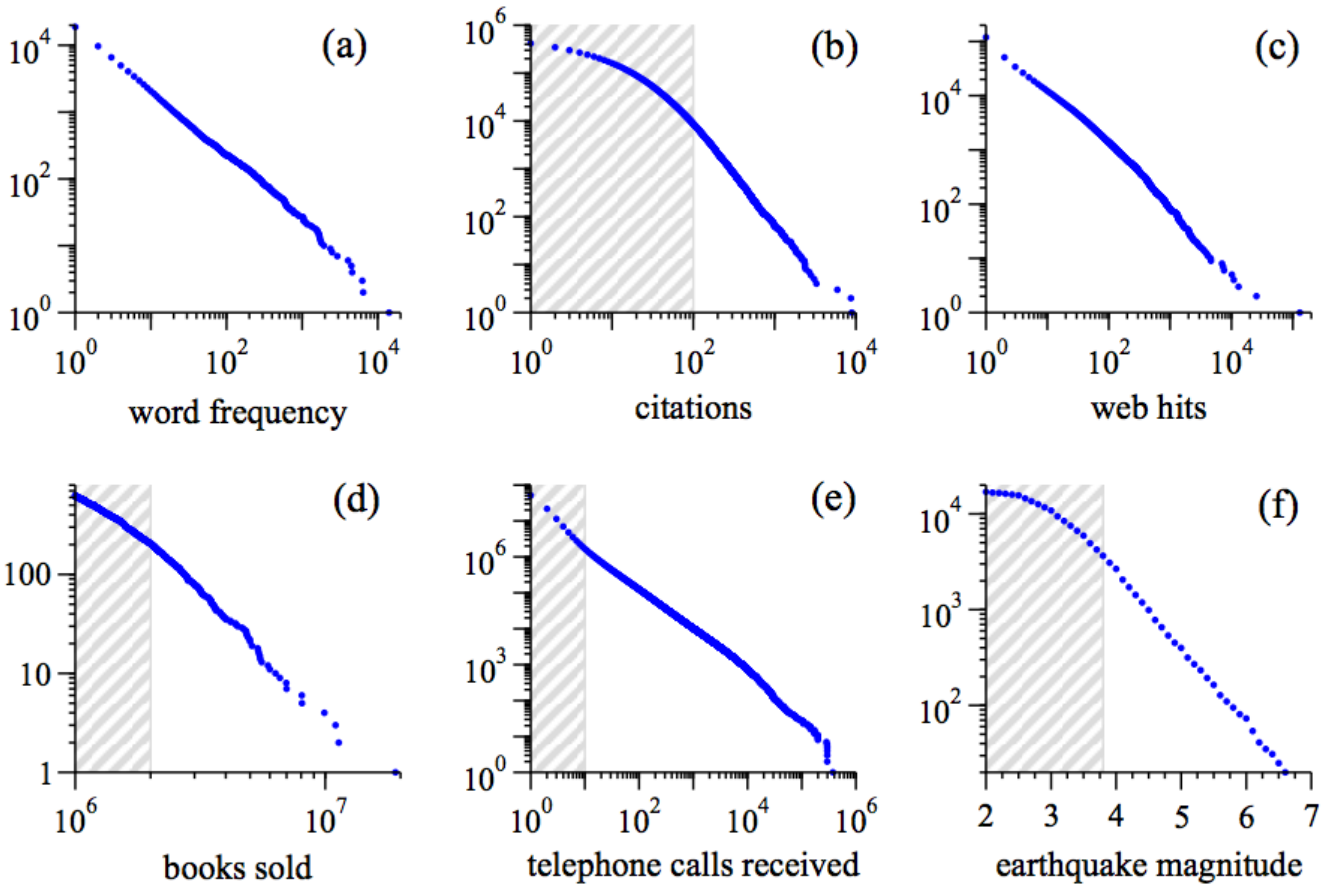
How to estimate the exponent α from samples?

- least squares fit is not reliable. e.g. least squares fit to plot (b) gives $\hat{\alpha}_{LS} = 2.26 \pm 0.02$.
- formula obtained as MLE:

$$\alpha = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]$$

Error can be derived by bootstrap/jackknife.
MLE estimate of simulated data $\hat{\alpha}_{MLE} = 2.500 \pm 0.002$

Examples of power laws



power law: lower cutoff $x_{\min}$

Remark:

- Real-world distributions typically follow power law only after some minimum value $x_{\min}$.
- One often hears a quantity "has a power law tail".
- A judgement is required to determine the value $x_{\min}$. One way is to perform a scan over all values of $x_{\min}$
- Once x_{min} is determined, the usual MLE estimate for α can be used.
- R package: poweRlaw

MLE Esimation: proof

$$
\begin{aligned}
p(x) = Cx^{-\alpha} &= \frac{\alpha - 1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-\alpha} \\
\mathcal{L} &= \ln \prod_{i=1}^{n} p(x_i) \\
&= \ln \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}}\left(\frac{x_i}{x_{\min}}\right)^{-\alpha} \\
&= n \ln(\alpha - 1) - n \ln x_{\min} - \alpha \sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}
\end{aligned}
\tag{1}
$$

Setting $\frac{\partial \mathcal{L}}{\partial \alpha} = 0$, we have

$$
\alpha = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]
$$

Zipf's law: an empirical power law for rank and frequency(size)

- Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.
- Let N be the number of elements, k be their rank, s be the value of the exponent characterizing the distribution.
  Zipf's law predicts that out of a population of N elements, the frequency of elements of rank k, $f(k; s, N)$, is:

$$f(k; s, N) = \frac{[constant]}{k^s}$$

- Pareto distribution and Zipf's law differ from each other in the way the C.D.F. is plotted. Unlike Pareto, Zipf's made the rank on x-axis and frequency on y-axis.

Zipf's law: Zipf's plot

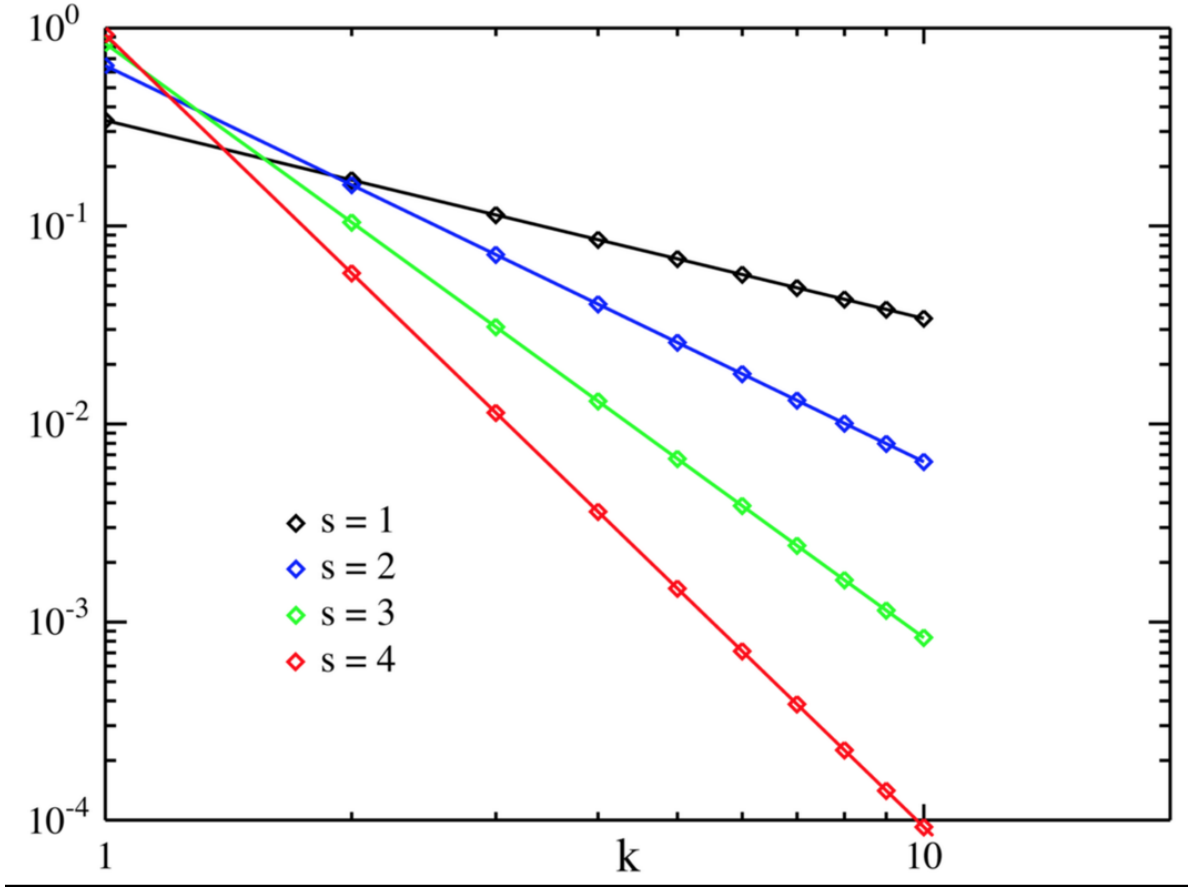


Figure : Zipf CDF for N = 10. The horizontal axis is the index k . (Note that the function is only defined at integer values of k. The connecting lines do not indicate continuity.)

Zipf's law: Estimation

Let the p.m.f of X with a Zipf's law be

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})}, x \geq x_{\min}$$

where $\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha}$

MLE esitmator numerically maximizes

$$\mathcal{L} = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln x_i$$

Details see Reference [3].

Zipf-Mandelbrot distribution: a generalization

A generalization of Zipf's law is the Zipf-Mandelbrot law, proposed by Benoit Mandelbrot, whose frequencies are:

$$f(k; N, q, s) = \frac{[constant]}{(k+q)^s}$$

Zipf's law: Explanation 1, optimization model

- Mandelbrot experiment: design a language over an alphabet of size d to optimize information per character.
  - Probability of jth most frequently used word is p_j
  - Length of jth most frequently used word is $c_j \sim \log_d j$
- Average information per word(entropy): $H = -\Sigma p_j \log p_j$
- Average characters per word: $C = \Sigma p_j c_j$
- Optimization leads to Zipf's law.

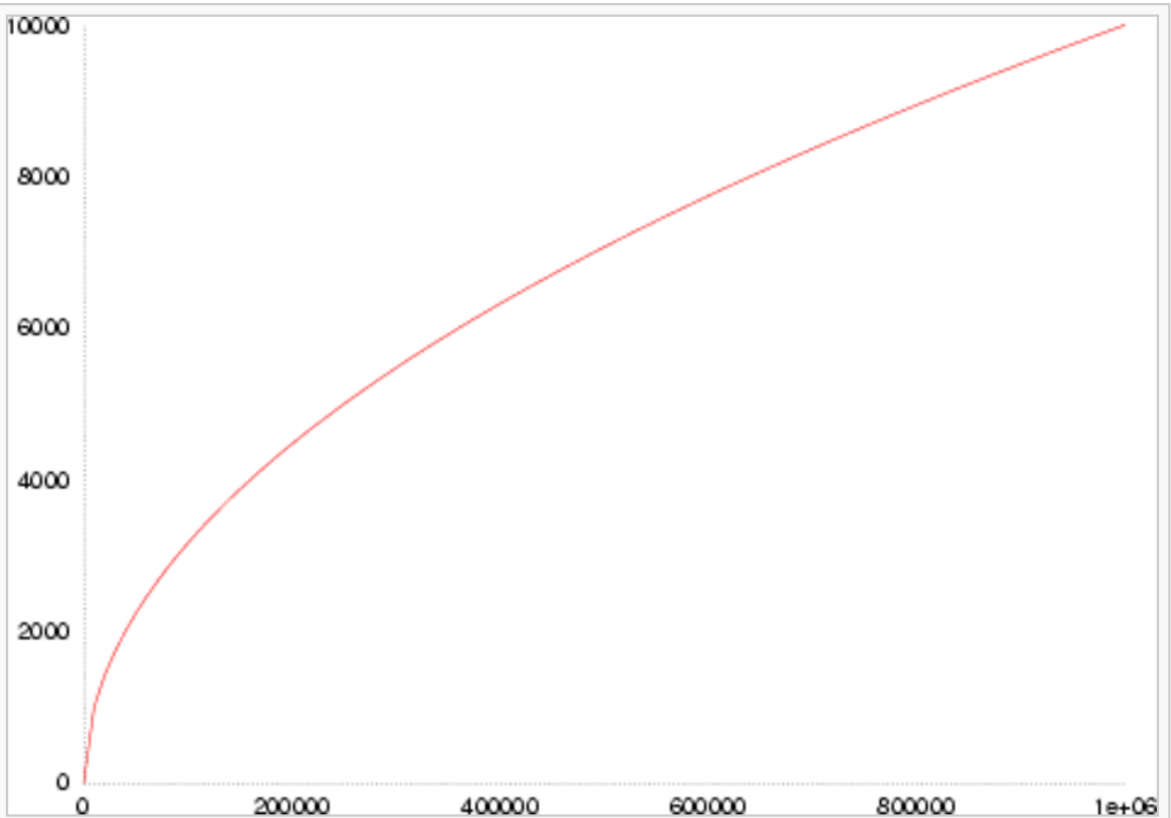
Zipf's law: Explanation 2, monkey typing randomly

- Miller(psychologist, 1957) suggests following: monkeys type randomly at a keyboard.
  - Hit each of n characters with probability p.
  - Hit space with probability $1 - np > 0$.
  - A word is a sequence of characters separated by a space
- Resulting distribution of word frequencies follows a Zipf's law.
- Conclusion: Mandelbrot's "optimization" not required for languages to have power law.

(optional) Preferential attachment

- Consider a dynamic Web graph.
  - pages join one at a time.
  - Each page has one outlink.
- Let $X_j(t)$ be the number of pages with degree j at time t.
- New page links:
  - With probability α, link to a random page.
  - With probability $1 - \alpha$, link to a page chosen proportionally to indegree.
- The resulting node degree follows a power law.

Heap's law



A typical Heaps-law plot. The x-axis represents the text size, and the y-axis represents the number of distinct vocabulary elements present in the text. Compare the values of the two axes

Heap's law

V_R: number of distinct words in an instanct text of size n.
K and β are free parameters.

$$V_R(n) = Kn^{\beta}$$

- can be derived from Zipf's law(asymptotically equivalent to Zipf's law under mild assumptions)
- implies diminishing returns in terms of discovery of the full vocabulary from which the distinct terms are drawn.
- is generalizable. E.g. objects - people, type - country

True or fake?

Table 1. One of the columns gives the land area of political states and territories in km^2. The other column contains faked data, generated with a random number generator.

State/Territory	Real or Faked Area (km^2)	
Afghanistan	645,807	796,467
Albania	28,748	9,943
Algeria	2,381,741	3,168,262
American Samoa	197	301
Andorra	464	577
Anguilla	96	82
Antigua and Barbuda	442	949
Argentina	2,777,409	4,021,545
Armenia	29,743	54,159
Aruba	193	367
Australia	7,682,557	6,563,132
Austria	83,858	64,154
Azerbaijan	86,530	71,661
Bahamas	13,962	9,125
Bahrain	694	755
Bangladesh	142,615	347,722
Barbados	431	818
Belgium	30,518	47,123
Belize	22,965	20,648
Benin	112,620	97,768
...	...	...

Benford's law (first digit law)

Imagine a large dataset, say something like a list of every country and its population.

COUNTRY	POPULATION
Afghanistan	**2**9,117,000
Albania	**3**,195,000
Algeria	**3**5,423,000
Andorra	**8**4,082
Angola	**1**8,993,000
	↑ Leading digit

Chances are, the leading digit will be a 1 more often than a 2. And 2s would probably occur more often than 3s, and so on.
This odd phenomenon is Benford's law.

Benford's law: mathematical statement

A set of numbers is said to satisfy Benford's law if the leading digit d ($d \in \{1, ..., 9\}$) occurs with probability

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\frac{d+1}{d}$$

d	*P*(*d*)	Relative size of *P*(*d*)
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

Figure : The distribution of first digits depicted by Benford's law.

▸ Testing Benford's Law

Benford's law: Basic Mechanism

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\frac{d+1}{d}$$

- Benford's law is expected if the mantissa(fractional part) of the **logarithms** of the numbers (but not the numbers themselves) are uniformly and randomly distributed.
- It tends to apply most accurately to data that are distributed uniformly across **many orders of magnitude**.

Benford's law: Possible Explanations

Examples:

- bacteria size: Outcomes of exponential growth processes. (e.g. 1000×2^n on day n)
  - exponentially growing quantity moves on a log-scale at a constant rate.
- stock price: Multiplicative fluctuations
  - logarithm of the stock price is undergoing a random walk
- The leading digits of data satisfying Zipf's law with $s = 1$ satisfies Benford's law.

Two major steps:

- scale invariance (Pinkham, 1961)
- mixture result (Ted Hill, 1995)

Benford's law:original table

Benford(1938) "collected data from as many fields as possible and to include a wide variety of types".

TABLE I

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Benford's law: Mixture result by Ted Hill, 1995

If one repeatedly "randomly" chooses a probability distribution (with some regularity conditions) and then randomly chooses numbers according to that distribution, the resulting list of numbers will obey Benford's Law.

Reference: Hill, Theodore P. "A statistical derivation of the significant-digit law." Statistical Science (1995): 354-363.

References

1. Wikipedia(Zipf's law, Heap's law, Benford's law)
2. Newman, Mark EJ. "Power laws, Pareto distributions and Zipf's law." Contemporary physics 46.5 (2005): 323-351.
3. Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." SIAM review 51.4 (2009): 661-703.
4. https://terrytao.wordpress.com/2009/07/03/benfords-law-zipfs-law-and-the-pareto-distribution/