Zipf's law (/ˈziːf/, not /ˈtзɪp/ as in German) is an empirical law formulated using mathematical statistics that refers to the fact that for many types of data studied in the physical and social sciences, the rank-frequency distribution is an inverse relation. The Zipfian distribution is one of a family of related discrete power law probability distributions. It is related to the zeta distribution, but is not identical.

Zipf's law was originally formulated in terms of quantitative linguistics, stating that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. For example, in the Brown Corpus of American English text, the word "the" is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences (69,971 out of slightly over 1 million). True to Zipf's Law, the second-place word "of" accounts for slightly over 3.5% of words (36,411 occurrences), followed by "and" (28,852). Only 135 vocabulary items are needed to account for half the Brown Corpus.[1]

The law is named after the American
linguist George Kingsley Zipf (1902–1950), who popularized it and sought to explain it (Zipf 1935, 1949), though he did not claim to have originated it.[2] The French stenographer Jean-Baptiste Estoup (1868–1950) appears to have noticed the regularity before Zipf.[3] It was also noted in 1913 by German physicist Felix Auerbach (1856–1933).[4]

The law is similar in concept, though not identical in distribution, to Benford's law.

| Parameters | $s \geq 0$ (real) |
| Support    | $N \in \{1, 2, 3 \ldots \}$ (integer) |
| PMF        | $\frac{1}{k^s} \quad \text{where } H_{N,s}$ is the $N$th generalized harmonic number |
| CDF        | $\frac{H_{k,s}}{H_{N,s}}$ |
| Mean       | $\frac{H_{N,s-1}}{H_{N,s}}$ |
| Mode       | 1 |
| Variance   | $\frac{H_{N,s-2}}{H_{N,s}} - \frac{H_{N,s-1}^2}{H_{N,s}^2}$ |
| Entropy    | $\frac{s}{H_{N,s}} \sum_{k=1}^{N} \ln(k) + \ln(H_{N,s})$ |
| MGF        | $\frac{1}{H_{N,s}} \sum_{n=1}^{N} e^{nt} n^s$ |
| CF         | $\frac{1}{H_{N,s}} \sum_{n=1}^{N} e^{int} n^s$ |

### Contents

- Other data sets
- Theoretical review
- Statistical explanation
- Mathematical explanation
- Related laws
- Applications
- See also
- References
- Further reading
- External links

### Other data sets

The same relationship occurs in many other rankings of human-created systems,[5] such as the ranks of mathematical expressions or ranks of notes in music,[6] and even in uncontrolled environments, such as corporation sizes, income rankings, ranks of number of people watching the same TV channel,[7] cells' transcriptomes[8] and so on. The appearance of the distribution in rankings of cities by population was first noticed by Felix Auerbach in 1913,[4] leading to a broad literature of Zipf's law for cities.[9] However, more recent empirical[10][11] and theoretical[12] studies have challenged the relevance of Zipf's law for cities.
Empirically, a data set can be tested to see whether Zipf’s law applies by checking the goodness of fit of an empirical distribution to the hypothesized power law distribution with a Kolmogorov–Smirnov test, and then comparing the (log) likelihood ratio of the power law distribution to alternative distributions like an exponential distribution or lognormal distribution.[13]

**Theoretical review**

Zipf’s law is most easily observed by plotting the data on a log-log graph, with the axes being log (rank order) and log (frequency). For example, the word “the” (as described above) would appear at $x = \log(1), y = \log(69971)$. It is also possible to plot reciprocal rank against frequency or reciprocal frequency or interword interval against rank.[2] The data conform to Zipf’s law to the extent that the plot is linear.

Formally, let:

- $N$ be the number of elements;
- $k$ be their rank;
- $s$ be the value of the exponent characterizing the distribution.

Zipf’s law then predicts that out of a population of $N$ elements, the normalized frequency of the element of rank $k$, $f(k; s, N)$, is:

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N} (1/n^s)}$$

Zipf’s law holds if the number of elements with a given frequency is a random variable with power law distribution $p(f) = \alpha f^{-1-1/s}$.[14]

It has been claimed that this representation of Zipf’s law is more suitable for statistical testing, and in this way it has been analyzed in more than 30,000 English texts. The goodness-of-fit tests yield that only about 15% of the texts are statistically compatible with this form of Zipf’s law. Slight variations in the definition of Zipf’s law can increase this percentage up to close to 50%.[15]

In the example of the frequency of words in the English language, $N$ is the number of words in the English language and, if we use the classic version of Zipf’s law, the exponent $s$ is 1. $f(k; s, N)$ will then be the fraction of the time the $k$th most common word occurs.

The law may also be written:

$$f(k; s, N) = \frac{1}{k^s H_{N,s}}$$
where \( H_{N,s} \) is the \( N \)th generalized harmonic number.

The simplest case of Zipf’s law is a "1/f" function. Given a set of Zipfian distributed frequencies, sorted from most common to least common, the second most common frequency will occur half as often as the first, the third most common frequency will occur \( \frac{1}{3} \) as often as the first, and the \( n \)th most common frequency will occur \( \frac{1}{n} \) as often as the first. However, this cannot hold exactly, because items must occur an integer number of times; there cannot be 2.5 occurrences of a word. Nevertheless, over fairly wide ranges, and to a fairly good approximation, many natural phenomena obey Zipf’s law.

In human languages, word frequencies have a very heavy-tailed distribution, and can therefore be modeled reasonably well by a Zipf distribution with an \( s \) close to 1.

As long as the exponent \( s \) exceeds 1, it is possible for such a law to hold with infinitely many words, since if \( s > 1 \) then

\[
\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} < \infty.
\]

where \( \zeta \) is Riemann’s zeta function.

### Statistical explanation

Although Zipf’s Law holds for all languages, even non-natural ones like Esperanto, the reason is still not well understood. However, it may be partially explained by the statistical analysis of randomly generated texts. Wentian Li has shown that in a document in which each character has been chosen randomly from a uniform distribution of all letters (plus a space character), the "words" with different lengths follow the macro-trend of the Zipf’s law (the more probable words are the shortest with equal probability). Vitold Belevitch, in a paper entitled *On the Statistical Laws of Linguistic Distribution*, offers a mathematical derivation. He took a large class of well-behaved statistical distributions (not only the normal distribution) and expressed them in terms of rank. He then expanded each expression into a Taylor series. In every case Belevitch obtained the remarkable result that a first-order truncation of the series resulted in Zipf’s law. Further, a second-order truncation of the Taylor series resulted in Mandelbrot’s law.

The principle of least effort is another possible explanation: Zipf himself proposed that neither speakers nor hearers using a given language want to work any harder than
necessary to reach understanding, and the process that results in approximately equal
distribution of effort leads to the observed Zipf distribution.[21][22]

Similarly, preferential attachment (intuitively, "the rich get richer" or "success breeds
success") that results in the Yule–Simon distribution has been shown to fit word frequency
versus rank in language[23] and population versus city rank[24] better than Zipf’s law. It was
originally derived to explain population versus rank in species by Yule, and applied to
cities by Simon.

Mathematical explanation

*Atlas models* are systems of exchangeable positive-valued *diffusion processes* with drift and
variance parameters that depend only on the rank of the process. It has been shown
mathematically that Zipf’s law holds for Atlas models that satisfy certain natural
regularity conditions.[25] Atlas models can be used to represent empirical systems of time-
dependent multivariate data, including, e.g., the frequency of words in a written language
or the size of companies. An Atlas model that represents an empirical system will have the
same stationary distribution as the empirical system, so if the Atlas model follows Zipf’s
law, the system will also follow Zipf’s law. Since Atlas models that satisfy natural
regularity conditions follow Zipf’s law, this accounts for its universality.[26]

In the figure above of the 10 million Wikipedia words, the log-log plots are not precisely
straight lines but rather slightly concave curves with a tangent of slope -1 at some point
along the curve. Such distributions are usually referred to as *quasi-Zipfian* distributions,
and most systems of time-dependent empirical data that are said to follow Zipf’s law are
actually quasi-Zipfian. Quasi-Zipfian systems can be represented by *quasi-Atlas models*,
and quasi-Atlas models are amenable to mathematical treatment similar to that for Zipf’s
law.

Related laws

*Zipf’s law* in fact refers more generally to frequency distributions of "rank data", in which
the relative frequency of the *n*th-ranked item is given by the zeta distribution, \(1/(n^s \zeta(s))\),
where the parameter \(s > 1\) indexes the members of this family of probability distributions.
Indeed, *Zipf’s law* is sometimes synonymous with "zeta distribution", since probability
distributions are sometimes called "laws". This distribution is sometimes called the
Zipfian distribution.

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

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

The "constant" is the reciprocal of the Hurwitz zeta function evaluated at *s*. In practice, as
easily observable in distribution plots for large corpora, the observed distribution can be modelled more accurately as a sum of separate distributions for different subsets or subtypes of words that follow different parameterizations of the Zipf–Mandelbrot distribution, in particular the closed class of functional words exhibit $s$ lower than 1, while open-ended vocabulary growth with document size and corpus size require $s$ greater than 1 for convergence of the Generalized Harmonic Series.[2]

Zipfian distributions can be obtained from Pareto distributions by an exchange of variables.[14]

The Zipf distribution is sometimes called the discrete Pareto distribution[27] because it is analogous to the continuous Pareto distribution in the same way that the discrete uniform distribution is analogous to the continuous uniform distribution.

The tail frequencies of the Yule–Simon distribution are approximately

$$f(k; \rho) \approx \frac{[\text{constant}]}{k^{\rho+1}}$$

for any choice of $\rho > 0$.

In the parabolic fractal distribution, the logarithm of the frequency is a quadratic polynomial of the logarithm of the rank. This can markedly improve the fit over a simple power-law relationship.[28] Like fractal dimension, it is possible to calculate Zipf dimension, which is a useful parameter in the analysis of texts.[29]

It has been argued that Benford's law is a special bounded case of Zipf's law,[28] with the connection between these two laws being explained by their both originating from scale invariant functional relations from statistical physics and critical phenomena.[30] The ratios of probabilities in Benford's law are not constant. The leading digits of data satisfying Zipf's law with $s = 1$ satisfy Benford's law.
### Applications

In information theory, a symbol (event, signal) of probability $p$ contains $-\log_2(1/p)$ bits of information. Hence, Zipf’s law for natural numbers: $\Pr(x) \approx 1/x$ is equivalent with number $x$ containing $\log_2(x)$ bits of information. To add information from a symbol of probability $p$ into information already stored in a natural number $x$, we should go to $x'$ such that $\log_2(x') \approx \log_2(x) + \log_2(1/p)$, or equivalently $x' \approx x/p$. For instance, in standard binary system we would have $x' = 2x + s$, what is optimal for $\Pr(s = 0) = \Pr(s = 1) = 1/2$ probability distribution. Using $x' \approx x/p$ rule for a general probability distribution is the base of Asymmetric Numeral Systems family of entropy coding methods used in data compression, whose state distribution is also governed by Zipf’s law.

Zipf’s law has been used for extraction of parallel fragments of texts out of comparable corpora.[31] Zipf’s law has also been used by Laurance Doyle and others at the SETI Institute as part of the search for extraterrestrial intelligence.[32][33]

The Voynich Manuscript, which is a 15th-century codex, also falls in line with Zipf’s law, indicating that text is most likely not a hoax but rather written in an obscure language or cipher.[34][35]

### See also

- 1% rule (Internet culture)
- Benford’s law
- Bradford’s law
- Brevity law
- Demographic gravitation
Frequency list
- Gibrat's law
- Hapax legomenon
- Heaps' law
- King effect
- Lorenz curve
- Lotka's law
- Menzerath's law
- Pareto distribution
- Pareto principle, a.k.a. the "80–20 rule"
- Price's law
- Principle of least effort
- Rank-size distribution
- Stigler's law of eponymy
- Long tail

References


17. Léon Brillouin, La science et la théorie de l'information, 1959, réédité en 1988, traduction anglaise rééditée en 2004


34. #author.fullName}. "Mystery text's language-like patterns may be an elaborate hoax" (https://www.newscientist.com/article/2106915-mystery-texts-language-like-patterns-may-be-an-elaborate-hoax/). *New Scientist*. Retrieved 2022-02-25. {}


**Further reading**

Primary:


Secondary:

- Frans J. Van Droogenbroeck (2016), Handling the Zipf distribution in computerized authorship attribution (https://www.academia.edu/24147736/)
- Frans J. Van Droogenbroeck (2019), An essential rephrasing of the Zipf-Mandelbrot law to solve authorship attribution applications by Gaussian statistics (https://www.academia.edu/40029629)
- Axtell, Robert L; Zipf distribution of US firm sizes (http://www.science mag.org/content/293/5536/1818.short), Science, 293, 5536, 1818, 2001, American Association for the Advancement of Science
External links

- Distributions de type "fractal parabolique" dans la Nature (French, with English summary) (http://www.hubbertpeak.com/laherrere/fractal.htm)
- An analysis of income distribution (https://www.newscientist.com/article.ns?id=mg18524904.300)
- Zipf List of French words (http://www.lexique.org/listes/liste_mots.txt)
- Zipf list for English, French, Spanish, Italian, Swedish, Icelandic, Latin, Portuguese and Finnish from Gutenberg Project and online calculator to rank words in texts (http://1.1o1.in/en/webtools/semantic-depth)
- Zipf's Law examples and modelling (1985) (http://www.geoffkirby.co.uk/ZIPFSLAW.pdf)