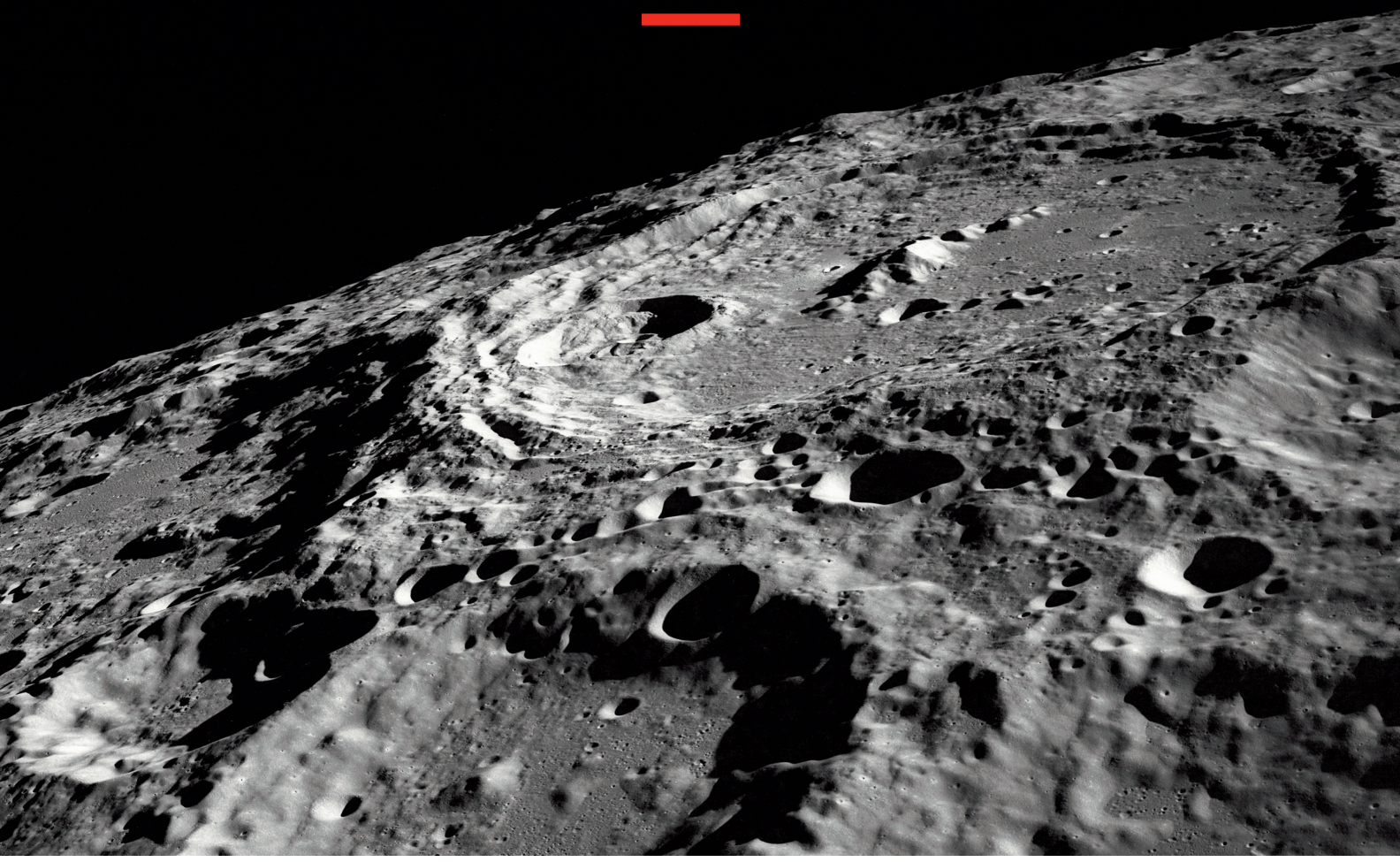
NOTEBOOK

Power-law distribution

Mark Newman describes a statistical distribution to represent data spanning many orders of magnitude, from the frequencies of words in human language to the sizes of craters on the moon



What is the power-law distribution?

Most quantities we deal with in statistics show clear clustering around an average value. The typical heights of adult human beings, for example, fall in a range from about 1 to 3 metres, with a mean somewhere around 2 meters. Such quantities are well fitted by familiar distributions like the bell-shaped normal distribution.

Some other quantities, however, depart from this pattern, with values that span many orders of magnitude and cannot, by any stretch of the imagination, be represented by a normal distribution. Take sales of books, for example. Many books are published each year that fail to gain traction and sell only a few copies. At the other end of the scale, the very best of the bestsellers sell tens of millions of copies, so the range from smallest observation to largest spans a factor of a million or more. Other examples of quantities with broad distributions include the populations of cities, the frequencies of words in human language, the numbers of "hits" on web pages, the numbers of citations of academic papers, the financial net worth of individuals, the magnitudes of earthquakes and solar flares, and the sizes of craters on the moon.

Mark Newman is the Anatol Rapoport Distinguished University Professor of Physics in the Department of Physics and the Center for the Study of Complex Systems, University of Michigan.

There are several statistical distributions used to represent quantities like these. One is the log-normal distribution,[1] but perhaps the most widely used is the power law. A power law is a statistical distribution over a quantity x following the form $x^{-\alpha}$, with α constant. Among other things, this means that when plotted on doubly logarithmic scales, the cumulative distribution function for a power law follows a straight line with slope $1 - \alpha$. Figure 1 shows an example. Such log-log plots are widely used as a

simple visual diagnostic for power-law behaviour, and measurement of the slope can give a rough estimate of α. One should be cautious, however: a simple straight-line fit to such a plot is known to give a biased estimate, and α is more properly calculated using maximum-likelihood methods.[2–4]

In most real-world applications, the value of α falls in the range $2 < \alpha \leq 3$, although values outside this range do occur on occasion.[5] If $\alpha \leq 2$ then the first moment of the distribution – its mean – is infinite, which is unrealistic in most cases. If $2 < \alpha \leq 3$ then the mean is finite but the second moment diverges, meaning the distribution has infinite variance. This behaviour has important repercussions, for instance in the theory of critical phenomena in physical systems and in models of the spread of epidemic disease.

Who discovered it?

The first well-known study is that of Vilfredo Pareto in 1896, who was interested in the distribution in the context of personal income.[6] The power law is sometimes called the Pareto distribution in his honour. It is also sometimes called Zipf's law, after George Zipf, who published another influential study in 1949.[7]

When should it be used?

The power law is one of several distributions used to represent positive-definite data with broad range, spanning many orders of magnitude. There are two situations in which power-law distributions are used. The first and more common of the two is driven by empirical observation: there are many cases where it is observed that a quantity approximately follows a power law, even though there may be no formal theory to explain why it does. Such behaviour is often first detected using a log-log plot of the type shown in Figure 1 and can then be evaluated quantitatively using rigorous statistical methods.[4] The power-law distribution is often taken as a hint at potential mechanisms underlying the observed data – there are many examples in the literature of the observation of a power law informing mechanistic theories in economics, physics, social science, biology, and other areas.

The second situation in which it is appropriate to use a power-law distribution is when there is a known mechanism at work that generates such distributions. Several such mechanisms exist, including the so-called Yule process (also known as preferential attachment or cumulative advantage), critical processes and self-organised criticality, and random walks.[5,8] An example of a case where the underlying mechanism is known is critical fluctuation in magnetic or fluid systems in physics.[9]

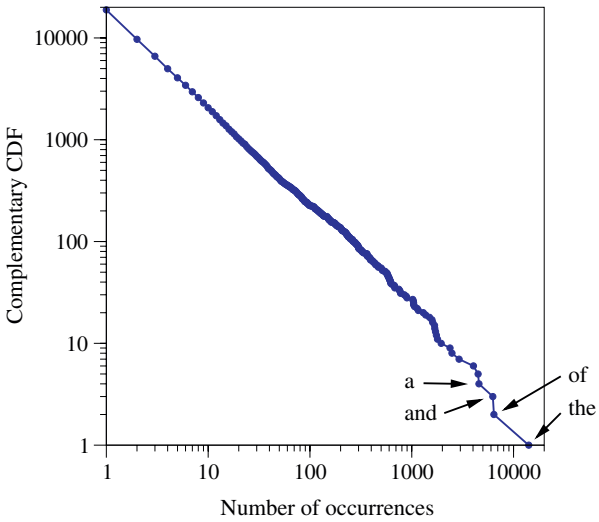


FIGURE 1 The complementary cumulative distribution function (CDF) of the frequencies of occurrence of English words in the novel *Moby Dick* by Herman Melville. The horizontal axis is the total number of times a word occurs in the book, and the vertical axis is the number of words which occur that many times or more. The four most common words – "the", "of", "and", and "a" – are represented by the last four data points on the right, as indicated. For a quantity obeying a power-law distribution, the complementary CDF, when plotted on doubly logarithmic scales as shown here, should follow a straight line, except for statistical fluctuations. The *Moby Dick* data appear to fit a power-law distribution well.[4]

When should it not be used?

Because use of the power-law distribution is most often motivated by observation, one can rarely be entirely certain of its correctness. The best one can say is that it is not ruled out by the data. There are, moreover, several alternative distributions, such as the log-normal or the stretched exponential, that can easily be confused with a power law. It is unfortunately quite common in the published literature for authors to rely solely on qualitative inspection of a log-log plot to detect power-law behaviour, and this is a not a reliable approach. A Kolmogorov–Smirnov or other nonparametric goodness-of-fit test is a better test for power-law behaviour.[4] Likelihood ratio tests can also be used to tell which of two hypothesised distributions is favoured.[10]

Keep in mind...

The power-law distribution is most often justified by empirical observation, not theory. Its observation can provide not only a compact representation of the distribution of certain quantities but also a hint as to the underlying mechanisms governing them. Log-log plots provide a simple tool for initial identification of possible power-law candidates, but more advanced statistical tools are required for rigorous testing of power-law behaviour. ■

A note about Notebook
This article is the third in our regular series on statistical distributions. Which do you want to read about next? Send suggestions – or a submission of your own – to significance@rss.org.uk.

References

1. Limpert, E. and Stahel, W. A. (2017) The log-normal distribution. *Significance*, **14**(1), 8–9.

2. Muniruzzaman, N. M. (1957) On measures of location and dispersion and tests of hypotheses in a Pareto population. *Bulletin of the Calcutta Statistical Association*, **7**, 115–123.

3. Hill, M. (1975) A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, **3**, 1163–1174.

4. Clauset, A., Shalizi, C. R. and Newman, M. E. J. (2009) Power-law distributions in empirical data. *SIAM Review*, **51**, 661–703.

5. Newman, M. E. J. (2005) Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, **46**, 323–351.

6. Pareto, V. (1896) *Cours d'Economie Politique*. Lausanne: F. Rouge.

7. Zipf, G. K. (1949) *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

8. Mitzenmacher, M. (2004) A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* **1**, 226–251.

9. Yeomans, J. M. (1992) *Statistical Mechanics of Phase Transitions*. Oxford: Oxford University Press.

10. Vuong, Q. H. (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307–333.

Main image: NASA