

Al-Kashi Background Statistics

<http://www.ar-php.org/stats/al-kashi/>

Contents

Articles

Summary Statistics	1
Mean	1
Median	6
Mode (statistics)	15
Variance	20
Standard deviation	32
Coefficient of variation	46
Skewness	49
Kurtosis	55
Ranking	61
Graphics	65
Box plot	65
Histogram	68
Q–Q plot	74
Ternary plot	79
Distributions	84
Normal distribution	84
Student's t-distribution	113
F-distribution	128
Feature scaling	131
Correlation and Regression	134
Covariance	134
Correlation and dependence	138
Regression analysis	144
Path analysis (statistics)	154
Analysis	156
Moving average	156
Student's t-test	162
Contingency table	171
Analysis of variance	174

Principal component analysis	189
Diversity index	205
Diversity index	205
Clustering	210
Hierarchical clustering	210
K-means clustering	215
Matrix	225
Matrix (mathematics)	225
Matrix addition	247
Matrix multiplication	249
Transpose	263
Determinant	266
Minor (linear algebra)	284
Adjugate matrix	287
Invertible matrix	291
Eigenvalues and eigenvectors	298
System of linear equations	316
References	
Article Sources and Contributors	326
Image Sources, Licenses and Contributors	332
Article Licenses	
License	335

Summary Statistics

Mean

In mathematics, **mean** has several different definitions depending on the context.

In probability and statistics, **mean** and expected value are used synonymously to refer to one measure of the central tendency either of a probability distribution or of the random variable characterized by that distribution. In the case of a discrete probability distribution of a random variable X , the mean is equal to the sum over every possible value weighted by the probability of that value; that is, it is computed by taking the product of each possible value x of X and its probability $P(x)$, and then adding all these products together, giving $\mu = \sum xP(x)$.^[1] An analogous formula applies to the case of a continuous probability distribution. Not every probability distribution has a defined mean; see the Cauchy distribution for an example. Moreover, for some distributions the mean is infinite: for example, when the probability of the value 2^n is $\frac{1}{2^n}$ for $n = 1, 2, 3, \dots$

For a data set, the terms arithmetic mean, mathematical expectation, and sometimes average are used synonymously to refer to a central value of a discrete set of numbers: specifically, the sum of the values divided by the number of values. The arithmetic mean of a set of numbers x_1, x_2, \dots, x_n is typically denoted by \bar{x} , pronounced "x bar". If the data set were based on a series of observations obtained by sampling from a statistical population, the arithmetic mean is termed the **sample mean** (denoted \bar{x}) to distinguish it from the **population mean** (denoted μ or μ_x).^[2]

For a finite population, the **population mean** of a property is equal to the arithmetic mean of the given property while considering every member of the population. For example, the population mean height is equal to the sum of the heights of every individual divided by the total number of individuals. The sample mean may differ from the population mean, especially for small samples. The law of large numbers dictates that the larger the size of the sample, the more likely it is that the sample mean will be close to the population mean.^[3]

Outside of probability and statistics, a wide range of other notions of "mean" are often used in geometry and analysis; examples are given below.

Types of mean

Pythagorean means

Arithmetic mean (AM)

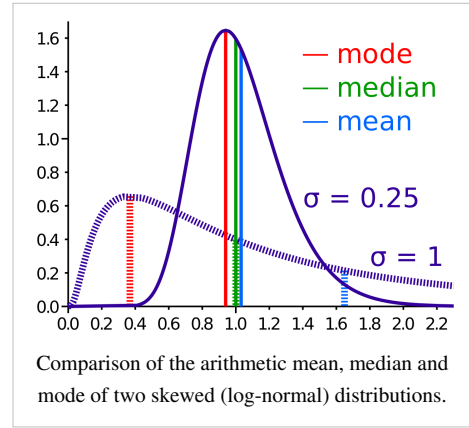
The *arithmetic mean* (or simply "mean") of a sample x_1, x_2, \dots, x_n is the sum the sampled values divided by the number of items in the sample:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

For example, the arithmetic mean of five values: 4, 36, 45, 50, 75 is

$$\frac{4 + 36 + 45 + 50 + 75}{5} = \frac{210}{5} = 42.$$

The **mean** may often be confused with the median, mode or range. The mean is the arithmetic average of a set of values, or distribution; however, for skewed distributions, the mean is not necessarily the same as the middle value (median), or the most likely (mode). For example, mean income is skewed upwards by a small number of people with very large incomes, so that the majority have an income lower than the mean. By contrast, the median income is the level at which half the population is below and half is above. The mode income is the most likely income, and favors the larger number of people with lower incomes. The median or mode are often more intuitive measures of such data.



Nevertheless, many skewed distributions are best described by their mean – such as the exponential and Poisson distributions.

Geometric mean (GM)

The geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean) e.g. rates of growth.

$$\bar{x} = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}$$

For example, the geometric mean of five values: 4, 36, 45, 50, 75 is:

$$(4 \times 36 \times 45 \times 50 \times 75)^{1/5} = \sqrt[5]{24\,300\,000} = 30.$$

Harmonic mean (HM)

The harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, for example speed (distance per unit of time).

$$\bar{x} = n \cdot \left(\sum_{i=1}^n \frac{1}{x_i} \right)^{-1}$$

For example, the harmonic mean of the five values: 4, 36, 45, 50, 75 is

$$\frac{5}{\frac{1}{4} + \frac{1}{36} + \frac{1}{45} + \frac{1}{50} + \frac{1}{75}} = \frac{5}{\frac{1}{3}} = 15.$$

Relationship between AM, GM, and HM

AM, GM, and HM satisfy these inequalities:

$$AM \geq GM \geq HM$$

Equality holds only when all the elements of the given sample are equal.

Generalized means

Power mean

The generalized mean, also known as the power mean or Hölder mean, is an abstraction of the quadratic, arithmetic, geometric and harmonic means. It is defined for a set of n positive numbers x_1 by

$$\bar{x}(m) = \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i^m \right)^{\frac{1}{m}}$$

By choosing different values for the parameter m , the following types of means are obtained:

$m \rightarrow \infty$	maximum
$m = 2$	quadratic mean
$m = 1$	arithmetic mean
$m \rightarrow 0$	geometric mean
$m = -1$	harmonic mean
$m \rightarrow -\infty$	minimum

***f*-mean**

This can be generalized further as the generalized f -mean

$$\bar{x} = f^{-1} \left(\frac{1}{n} \cdot \sum_{i=1}^n f(x_i) \right)$$

and again a suitable choice of an invertible f will give

$f(x) = x$	arithmetic mean,
$f(x) = \frac{1}{x}$	harmonic mean,
$f(x) = x^m$	power mean,
$f(x) = \ln x$	geometric mean.

Weighted arithmetic mean

The weighted arithmetic mean (or weighted average) is used if one wants to combine average values from samples of the same population with different sample sizes:

$$\bar{x} = \frac{\sum_{i=1}^n w_i \cdot x_i}{\sum_{i=1}^n w_i}.$$

The weights w_i represent the sizes of the different samples. In other applications they represent a measure for the reliability of the influence upon the mean by the respective values.

Truncated mean

Sometimes a set of numbers might contain outliers, i.e., data values which are much lower or much higher than the others. Often, outliers are erroneous data caused by artifacts. In this case, one can use a truncated mean. It involves discarding given parts of the data at the top or the bottom end, typically an equal amount at each end, and then taking the arithmetic mean of the remaining data. The number of values removed is indicated as a percentage of total number of values.

Interquartile mean

The interquartile mean is a specific example of a truncated mean. It is simply the arithmetic mean after removing the lowest and the highest quarter of values.

$$\bar{x} = \frac{2}{n} \sum_{i=(n/4)+1}^{3n/4} x_i$$

assuming the values have been ordered, so is simply a specific example of a weighted mean for a specific set of weights.

Mean of a function

In calculus, and especially multivariable calculus, the mean of a function is loosely defined as the average value of the function over its domain. In one variable, the mean of a function $f(x)$ over the interval (a, b) is defined by

$$\bar{f} = \frac{1}{b-a} \int_a^b f(x) dx.$$

Recall that a defining property of the average value \bar{y} of finitely many numbers y_1, y_2, \dots, y_n is that $n\bar{y} = y_1 + y_2 + \dots + y_n$. In other words, \bar{y} is the *constant* value which when *added* to itself n times equals the result of adding the n terms of y_i . By analogy, a defining property of the average value \bar{f} of a function over the interval $[a, b]$ is that

$$\int_a^b \bar{f} dx = \int_a^b f(x) dx$$

In other words, \bar{f} is the *constant* value which when *integrated* over $[a, b]$ equals the result of integrating $f(x)$ over $[a, b]$. But by the second fundamental theorem of calculus, the integral of a constant \bar{f} is just

$$\int_a^b \bar{f} dx = \bar{f}x \Big|_a^b = \bar{f}b - \bar{f}a = (b-a)\bar{f}$$

See also the first mean value theorem for integration, which guarantees that if f is continuous then there exists a point $c \in (a, b)$ such that

$$\int_a^b f(x) dx = f(c)(b-a)$$

The point $f(c)$ is called the mean value of $f(x)$ on $[a, b]$. So we write $\bar{f} = f(c)$ and rearrange the preceding equation to get the above definition.

In several variables, the mean over a relatively compact domain U in a Euclidean space is defined by

$$\bar{f} = \frac{1}{\text{Vol}(U)} \int_U f.$$

This generalizes the **arithmetic** mean. On the other hand, it is also possible to generalize the **geometric** mean to functions by defining the geometric mean of f to be

$$\exp\left(\frac{1}{\text{Vol}(U)} \int_U \log f\right).$$

More generally, in measure theory and probability theory, either sort of mean plays an important role. In this context, Jensen's inequality places sharp estimates on the relationship between these two different notions of the mean of a function.

There is also a *harmonic average* of functions and a *quadratic average* (or *root mean square*) of functions.

Mean of a probability distribution

See expected value.

Mean of angles

Sometimes the usual calculations of means fail on cyclical quantities such as angles, times of day, and other situations where modular arithmetic is used. For those quantities it might be appropriate to use a mean of circular quantities to take account of the modular values, or to adjust the values before calculating the mean.

Fréchet mean

The Fréchet mean gives a manner for determining the "center" of a mass distribution on a surface or, more generally, Riemannian manifold. Unlike many other means, the Fréchet mean is defined on a space whose elements cannot necessarily be added together or multiplied by scalars. It is sometimes also known as the **Karcher mean** (named after Hermann Karcher).

Other means

- Arithmetic-geometric mean
- Arithmetic-harmonic mean
- Cesàro mean
- Chisini mean
- Contraharmonic mean
- Distance-weighted estimator
- Elementary symmetric mean
- Geometric-harmonic mean
- Heinz mean
- Heronian mean
- Identric mean
- Lehmer mean
- Logarithmic mean
- Median
- Moving average
- Root mean square
- Rényi's entropy (a generalized f-mean)
- Stolarsky mean
- Weighted geometric mean
- Weighted harmonic mean

Distribution of the population mean

Using the sample mean

The arithmetic mean of a population, or population mean, is denoted μ . The sample mean (the arithmetic mean of a sample of values drawn from the population) makes a good estimator of the population mean, as its expected value is equal to the population mean (that is, it is an unbiased estimator). The sample mean is a random variable, not a constant, since its calculated value will randomly differ depending on which members of the population are sampled, and consequently it will have its own distribution. For a random sample of n observations from a normally distributed population, the sample mean distribution is normally distributed with mean and variance as follows:

$$\bar{x} \sim N \left\{ \mu, \frac{\sigma^2}{n} \right\}.$$

Often, since the population *variance* is an unknown parameter, it is estimated by the mean sum of squares; when this estimated value is used, the distribution of the sample mean is no longer a normal distribution but rather a Student's *t* distribution with $n - 1$ degrees of freedom.

References

- [1] Elementary Statistics by Robert R. Johnson and Patricia J. Kuby, p. 279 (<http://books.google.com/books?id=DWCAh7jWO98C&lpg=PP1&pg=PA279#v=onepage&q&f=false>)
- [2] Underhill, L.G.; Bradfield d. (1998) *Introstat*, Juta and Company Ltd. ISBN 0-7021-3838-X p. 181 (<http://books.google.com/books?id=f6TIVjrSAsG&lpg=PP1&pg=PA181#v=onepage&q&f=false>)
- [3] Schaum's Outline of Theory and Problems of Probability by Seymour Lipschutz and Marc Lipson, p. 141 (<http://books.google.com/books?id=ZKdqlw2ZnAMC&lpg=PP1&pg=PA141#v=onepage&q&f=false>)

External links

- Weisstein, Eric W., " Mean (<http://mathworld.wolfram.com/Mean.html>)", *MathWorld*.
- Weisstein, Eric W., " Arithmetic Mean (<http://mathworld.wolfram.com/ArithmeticMean.html>)", *MathWorld*.
- Comparison between arithmetic and geometric mean of two numbers (<http://www.sengpielaudio.com/calculator-geommean.htm>)
- Some relationships involving means (<http://www.math.uni-bielefeld.de/~sillke/PUZZLES/means-trapezoid>)

Median

In statistics and probability theory, the **median** is the numerical value separating the higher half of a data sample, a population, or a probability distribution, from the lower half. The *median* of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one (e.g., the median of {3, 3, 5, 9, 11} is 5). If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values ^[1] (the median of {3, 5, 7, 9} is $(5 + 7) / 2 = 6$), which corresponds to interpreting the median as the fully trimmed mid-range. The median is of central importance in robust statistics, as it is the most resistant statistic, having a breakdown point of 50%: so long as no more than half the data is contaminated, the median will not give an arbitrarily large result. A median is only defined on ordered one-dimensional data, and is independent of any distance metric. A geometric median, on the other hand, is defined in any number of dimensions.

In a sample of data, or a finite population, there may be no member of the sample whose value is identical to the median (in the case of an even sample size); if there is such a member, there may be more than one so that the median may not uniquely identify a sample member. Nonetheless, the value of the median is uniquely determined with the usual definition. A related concept, in which the outcome is forced to correspond to a member of the sample, is the medoid. At most, half the population have values strictly less than the *median*, and, at most, half have values strictly greater than the median. If each group contains less than half the population, then some of the population is exactly equal to the median. For example, if $a < b < c$, then the median of the list $\{a, b, c\}$ is b , and, if $a < b < c < d$, then the median of the list $\{a, b, c, d\}$ is the mean of b and c ; i.e., it is $(b + c)/2$.

The median can be used as a measure of location when a distribution is skewed, when end-values are not known, or when one requires reduced importance to be attached to outliers, e.g., because they may be measurement errors.

In terms of notation, some authors represent the median of a variable x either as \tilde{x} or as $\mu_{1/2}$, sometimes also M . There is no widely accepted standard notation for the median, so the use of these or other symbols for the median needs to be explicitly defined when they are introduced.

The median is the 2nd quartile, 5th decile, and 50th percentile.

Measures of location and dispersion

The median is one of a number of ways of summarising the typical values associated with members of a statistical population; thus, it is a possible location parameter.

When the median is used as a location parameter in descriptive statistics, there are several choices for a measure of variability: the range, the interquartile range, the mean absolute deviation, and the median absolute deviation. Since the median is the same as the *second quartile*, its calculation is illustrated in the article on quartiles.

For practical purposes, different measures of location and dispersion are often compared on the basis of how well the corresponding population values can be estimated from a sample of data. The median, estimated using the sample median, has good properties in this regard. While it is not usually optimal if a given population distribution is assumed, its properties are always reasonably good. For example, a comparison of the efficiency of candidate estimators shows that the sample mean is more statistically efficient than the sample median when data are uncontaminated by data from heavy-tailed distributions or from mixtures of distributions, but less efficient otherwise, and that the efficiency of the sample median is higher than that for a wide range of distributions. More specifically, the median has a 64% efficiency compared to the minimum-variance mean (for large normal samples), which is to say the variance of the median will be ~50% greater than the variance of the mean—see Efficiency (statistics)#Asymptotic efficiency and references therein.

Medians of probability distributions

For any probability distribution on the real line \mathbf{R} with cumulative distribution function F , regardless of whether it is any kind of continuous probability distribution, in particular an absolutely continuous distribution (which has a probability density function), or a discrete probability distribution, a median is by definition any real number m that satisfies the inequalities

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

or, equivalently, the inequalities

$$\int_{(-\infty, m]} dF(x) \geq \frac{1}{2} \text{ and } \int_{[m, \infty)} dF(x) \geq \frac{1}{2}$$

in which a Lebesgue–Stieltjes integral is used. For an absolutely continuous probability distribution with probability density function f , the median satisfies

$$P(X \leq m) = P(X \geq m) = \int_{-\infty}^m f(x) dx = \frac{1}{2}.$$

Any probability distribution on \mathbf{R} has at least one median, but there may be more than one median. Where exactly one median exists, statisticians speak of "the median" correctly; even when the median is not unique, some statisticians speak of "the median" informally.

Medians of particular distributions

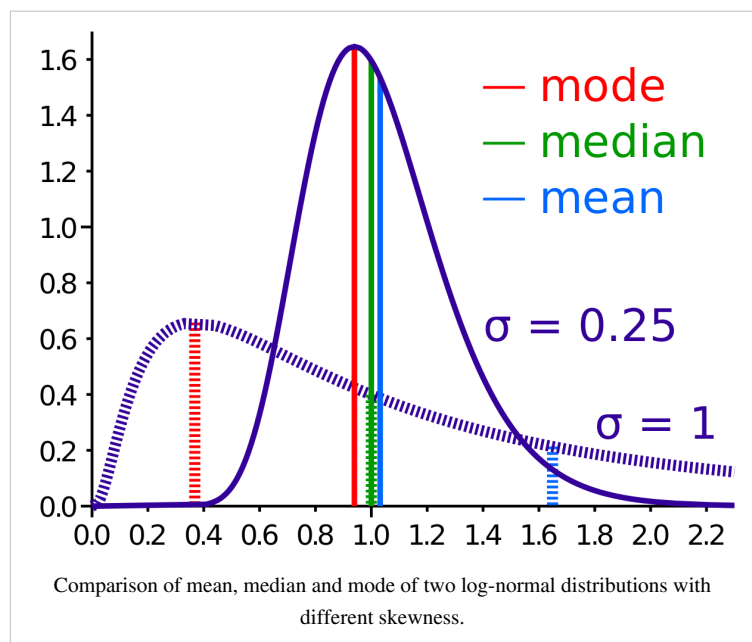
The medians of certain types of distributions can be easily calculated from their parameters:

- The median of a symmetric distribution with mean μ is μ .
 - The median of a normal distribution with mean μ and variance σ^2 is μ . In fact, for a normal distribution, mean = median = mode.
 - The median of a uniform distribution in the interval $[a, b]$ is $(a + b) / 2$, which is also the mean.
- The median of a Cauchy distribution with location parameter x_0 and scale parameter y is x_0 , the location parameter.
- The median of an exponential distribution with rate parameter λ is the natural logarithm of 2 divided by the rate parameter: $\lambda^{-1} \ln 2$.
- The median of a Weibull distribution with shape parameter k and scale parameter λ is $\lambda(\ln 2)^{1/k}$.

Medians in descriptive statistics

The median is used primarily for skewed distributions, which it summarizes differently than the arithmetic mean. Consider the multiset $\{ 1, 2, 2, 2, 3, 14 \}$. The median is 2 in this case, (as is the mode), and it might be seen as a better indication of central tendency (less susceptible to the exceptionally large value in data) than the arithmetic mean of 4.

Calculation of medians is a popular technique in summary statistics and summarizing statistical data, since it is simple to understand and easy to calculate, while also giving a measure that is more robust in the presence of outlier values than is the mean.



Medians for populations

An optimality property

The *mean absolute error* of a real variable c with respect to the random variable X is

$$E(|X - c|)$$

Provided that the probability distribution of X is such that the above expectation exists, then m is a median of X if and only if m is a minimizer of the mean absolute error with respect to X . In particular, m is a sample median if and only if m minimizes the arithmetic mean of the absolute deviations.

See also k -medians clustering.

Unimodal distributions

It can be shown for a unimodal distribution that the median \tilde{X} and the mean \bar{X} lie within $(3/5)^{1/2} \approx 0.7746$ standard deviations of each other.^[2] In symbols,

$$\frac{|\tilde{X} - \bar{X}|}{\sigma} \leq (3/5)^{1/2}$$

where $|\cdot|$ is the absolute value.

A similar relation holds between the median and the mode: they lie within $3^{1/2} \approx 1.732$ standard deviations of each other:

$$\frac{|\tilde{X} - \text{mode}|}{\sigma} \leq 3^{1/2}.$$

An inequality relating means and medians

If the distribution has finite variance, then the distance between the median and the mean is bounded by one standard deviation.

This bound was proved by Mallows, who used Jensen's inequality twice, as follows. We have

$$\begin{aligned} |\mu - m| &= |\mathbf{E}(X - m)| \leq \mathbf{E}(|X - m|) \\ &\leq \mathbf{E}(|X - \mu|) \\ &\leq \sqrt{\mathbf{E}((X - \mu)^2)} = \sigma. \end{aligned}$$

The first and third inequalities come from Jensen's inequality applied to the absolute-value function and the square function, which are each convex. The second inequality comes from the fact that a median minimizes the absolute deviation function

$$a \mapsto \mathbf{E}(|X - a|).$$

This proof can easily be generalized to obtain a multivariate version of the inequality, as follows:

$$\begin{aligned} \|\mu - m\| &= \|\mathbf{E}(X - m)\| \leq \mathbf{E}\|X - m\| \\ &\leq \mathbf{E}(\|X - \mu\|) \\ &\leq \sqrt{\mathbf{E}(\|X - \mu\|^2)} = \sqrt{\text{trace}(\text{var}(X))} \end{aligned}$$

where m is a spatial median, that is, a minimizer of the function $a \mapsto \mathbf{E}(\|X - a\|)$. The spatial median is unique when the data-set's dimension is two or more. An alternative proof uses the one-sided Chebyshev inequality; it appears in an inequality on location and scale parameters.

Jensen's inequality for medians

Jensen's inequality states that for any random variable x with a finite expectation $E(x)$ and for any convex function f

$$f(E(x)) \leq E(f(x))$$

It has been shown that if x is a real variable with a unique median m and f is a C function then

$$f(m) \leq \text{Median}(f(x))$$

A C function is a real valued function, defined on the set of real numbers R , with the property that for any real t

$$f^{-1}((-\infty, t]) = \{x \in R | f(x) \leq t\}$$

is a closed interval, a singleton or an empty set.

Medians for samples

The sample median

Efficient computation of the sample median

Even though comparison-sorting n items requires $\Omega(n \log n)$ operations, selection algorithms can compute the k^{th} -smallest of n items with only $\Theta(n)$ operations. This includes the median, which is the $\lfloor n/2 \rfloor$ th order statistic (or for an odd number of samples, the average of the two middle order statistics).

Easy explanation of the sample median

In individual series (if number of observation is very low) first one must arrange all the observations in ascending order. Then count(n) is the total number of observation in given data.

If n is odd then Median (M) = value of $((n + 1)/2)$ th item term.

If n is even then Median (M) = value of $[(n)/2]$ th item term + $((n)/2 + 1)$ th item term $] / 2$

For an odd number of values

As an example, we will calculate the sample median for the following set of observations: 1, 5, 2, 8, 7.

Start by sorting the values: 1, 2, 5, 7, 8.

In this case, the median is 5 since it is the middle observation in the ordered list.

The median is the $((n + 1)/2)$ th item, where n is the number of values. For example, for the list {1, 2, 5, 7, 8}, we have $n = 5$, so the median is the $((5 + 1)/2)$ th item.

$$\text{median} = (6/2)\text{th item}$$

$$\text{median} = 3\text{rd item}$$

$$\text{median} = 5$$

For an even number of values

As an example, we will calculate the sample median for the following set of observations: 1, 6, 2, 8, 7, 2.

Start by sorting the values: 1, 2, 2, 6, 7, 8.

In this case, the arithmetic mean of the two middlemost terms is $(2 + 6)/2 = 4$. Therefore, the median is 4 since it is the arithmetic mean of the middle observations in the ordered list.

We also use this formula $\text{MEDIAN} = \{(n + 1)/2\}$ th item . n = number of values

As above example 1, 2, 2, 6, 7, 8 $n = 6$ Median = $\{(6 + 1)/2\}$ th item = 3.5th item. In this case, the median is average of the 3rd number and the next one (the fourth number). The median is $(2 + 6)/2$ which is 4.

Variance

The distribution of both the sample mean and the sample median were determined by Laplace. The distribution of the sample median from a population with a density function $f(x)$ is asymptotically normal with mean m and variance

$$\frac{1}{4n f(m)^2}$$

where m is the median value of distribution and n is the sample size. In practice this may be difficult to estimate as the density function is usually unknown.

These results have also been extended. It is now known that for the p -th quartile that the distribution of the sample p -th quartile is distributed normally around the p -th quartile with variance equal to

$$\frac{p(1-p)}{nf(x_p)^2}$$

where $f(x_p)$ is the value of the distribution at the p -th quartile.

Estimation of variance from sample data

The value of $(2f(x))^{-2}$ —the asymptotic value of $n^{-\frac{1}{2}}(\nu - m)$ where ν is the population median—has been studied by several authors. The standard 'delete one' jackknife method produces inconsistent results. An alternative—the 'delete k' method—where k grows with the sample size has been shown to be asymptotically consistent. This method may be computationally expensive for large data sets. A bootstrap estimate is known to be consistent, but converges very slowly (order of $n^{-\frac{1}{4}}$). Other methods have been proposed but their behavior may differ between large and small samples.

Efficiency

The efficiency of the sample median, measured as the ratio of the variance of the mean to the variance of the median, depends on the sample size and on the underlying population distribution. For a sample of size $N = 2n + 1$ from the normal distribution, the ratio is

$$\frac{4n}{\pi(2n + 1)}$$

For large samples (as n tends to infinity) this ratio tends to $\frac{2}{\pi}$.

Other estimators

For univariate distributions that are *symmetric* about one median, the Hodges–Lehmann estimator is a robust and highly efficient estimator of the population median.

If data are represented by a statistical model specifying a particular family of probability distributions, then estimates of the median can be obtained by fitting that family of probability distributions to the data and calculating the theoretical median of the fitted distribution.^[citation needed] Pareto interpolation is an application of this when the population is assumed to have a Pareto distribution.

Coefficient of dispersion

The coefficient of dispersion (CD) is defined as the ratio of the average absolute deviation from the median to the median of the data.^[3] It is a statistical measure used by the states of Iowa, New York and South Dakota in estimating dues taxes.^{[4][5][6]} In symbols

$$CD = \frac{1}{n} \sum \frac{|m - x|}{m}$$

where n is the sample size, m is the sample median and x is a variate. The sum is taken over the whole sample.

Confidence intervals for a two sample test where the sample sizes are large have been derived by Bonett and Seier This test assumes that both samples have the same median but differ in the dispersion around it. The confidence interval (CI) is bounded inferiorly by

$$\exp \left[\log \left(\frac{t_a}{t_b} \right) - z_\alpha \left(\text{var} \left[\log \left(\frac{t_a}{t_b} \right) \right] \right)^{0.5} \right]$$

where t_j is the mean absolute deviation of the j^{th} sample, $\text{var}()$ is the variance and z_α is the value from the normal distribution for the chosen value of α : for $\alpha = 0.05$, $z_\alpha = 1.96$. The following formulae are used in the derivation of these confidence intervals

$$\text{var}[\log(t_a)] = \frac{\left(\frac{s_a^2}{t_a^2} + \left(\frac{x_a - \bar{x}}{t_a} \right)^2 - 1 \right)}{n}$$

$$\text{var}[\log(t_a/t_b)] = \text{var}[\log(t_a)] + \text{var}[\log(t_b)] - 2r(\text{var}[\log(t_a)]\text{var}[\log(t_b)])^{0.5}$$

where r is the Pearson correlation coefficient between the squared deviation scores

$$d_{ia} = |x_{ia} - \bar{x}_a| \text{ and } d_{ib} = |x_{ib} - \bar{x}_b|$$

a and b here are constants equal to 1 and 2, x is a variate and s is the standard deviation of the sample.

Multivariate median

Previously, this article discussed the concept of a univariate median for a one-dimensional object (population, sample). When the dimension is two or higher, there are multiple concepts that extend the definition of the univariate median; each such multivariate median agrees with the univariate median when the dimension is exactly one. In higher dimensions, however, there are several multivariate medians.

Marginal median

The marginal median is defined for vectors defined with respect to a fixed set of coordinates. A marginal median is defined to be the vector whose components are univariate medians. The marginal median is easy to compute, and its properties were studied by Puri and Sen.^[7]

Spatial median (L1 median)

In a normed vector space of dimension two or greater, the "spatial median" minimizes the expected distance

$$a \mapsto \mathbf{E}(\|X - a\|),$$

where X and a are vectors, if this expectation has a finite minimum; another definition is better suited for general probability-distributions. The spatial median is unique when the data-set's dimension is two or more. It is a robust and highly efficient estimator of the population spatial-median (also called the "L1 median"). Wikipedia:Please clarify

Other multivariate medians

An alternative to the spatial median is defined in a similar way, but based on a different loss function, and is called the Geometric median.^[citation needed] The centerpoint is another generalization to higher dimensions that does not relate to a particular metric.

Other median-related concepts

Pseudo-median

For univariate distributions that are *symmetric* about one median, the Hodges–Lehmann estimator is a robust and highly efficient estimator of the population median; for non-symmetric distributions, the Hodges–Lehmann estimator is a robust and highly efficient estimator of the population *pseudo-median*, which is the median of a symmetrized distribution and which is close to the population median.^[citation needed] The Hodges–Lehmann estimator has been generalized to multivariate distributions.

Variants of regression

The Theil–Sen estimator is a method for robust linear regression based on finding medians of slopes.^[citation needed]

Median filter

In the context of image processing of monochrome raster images there is a type of noise, known as the salt and pepper noise, when each pixel independently becomes black (with some small probability) or white (with some small probability), and is unchanged otherwise (with the probability close to 1). An image constructed of median values of neighborhoods (like 3×3 square) can effectively reduce noise in this case.^[citation needed]

Cluster analysis

In cluster analysis, the k-medians clustering algorithm provides a way of defining clusters, in which the criterion of maximising the distance between cluster-means that is used in k-means clustering, is replaced by maximising the distance between cluster-medians.

Median-Median Line

This is a method of robust regression. The idea dates back to Wald in 1940 who suggested dividing a set of bivariate data into two halves depending on the value of the independent parameter x : a left half with values less than the median and a right half with values greater than the median. He suggested taking the means of the dependent y and independent x variables of the left and the right halves and estimating the slope of the line joining these two points. The line could then be adjusted to fit the majority of the points in the data set.

Nair and Shrivastava in 1942 suggested a similar idea but instead advocated dividing the sample into three equal parts before calculating the means of the subsamples. Brown and Mood in 1951 proposed the idea of using the medians of two subsamples rather the means. Tukey combined these ideas and recommended dividing the sample into three equal size subsamples and estimating the line based on the medians of the subsamples.

Median-unbiased estimators

Any *mean*-unbiased estimator minimizes the risk (expected loss) with respect to the squared-error loss function, as observed by Gauss. A *median*-unbiased estimator minimizes the risk with respect to the absolute-deviation loss function, as observed by Laplace. Other loss functions are used in statistical theory, particularly in robust statistics.

The theory of median-unbiased estimators was revived by George W. Brown^[8] in 1947:

An estimate of a one-dimensional parameter θ will be said to be median-unbiased if, for fixed θ , the median of the distribution of the estimate is at the value θ ; i.e., the estimate underestimates just as often as it overestimates. This requirement seems for most purposes to accomplish as much as the mean-unbiased requirement and has the additional property that it is invariant under one-to-one transformation. [page 584]

Further properties of median-unbiased estimators have been reported. In particular, median-unbiased estimators exist in cases where mean-unbiased and maximum-likelihood estimators do not exist. Median-unbiased estimators are invariant under one-to-one transformations.

History

The idea of the median originated^[citation needed] in Edward Wright's book on navigation (*Certain Errors in Navigation*) in 1599 in a section concerning the determination of location with a compass. Wright felt that this value was the most likely to be the correct value in a series of observations.

In 1757, Roger Joseph Boscovich developed a regression method based on the L1 norm and therefore implicitly on the median.

The distribution of both the sample mean and the sample median were determined by Laplace in the early 1800s.^[9]

Antoine Augustin Cournot in 1843 was the first^[citation needed] to use the term *median* (*valeur médiane*) for the value that divides a probability distribution into two equal halves. Gustav Theodor Fechner used the median (*Centralwerth*) in sociological and psychological phenomena.^[10] It had earlier been used only in astronomy and related fields. Gustav Fechner popularized the median into the formal analysis of data, although it had been used previously by Laplace.

Francis Galton used the English term *median* in 1881,^[11] having earlier used the terms *middle-most value* in 1869 and the *medium* in 1880.^[citation needed]

References

- [1] http://www.stat.psu.edu/old_resources/ClassNotes/ljs_07/sld008.htm Simon, Laura J.; "Descriptive statistics", *Statistical Education Resource Kit*, Pennsylvania State Department of Statistics
- [2] <http://www.se16.info/hgb/cheb2.htm#3unimodalequalities>
- [3] Bonett DG, Seier E (2006) Confidence interval for a coefficient of dispersion in non-normal distributions. *Biometrical Journal* 48 (1) 144-148
- [4] http://www.iowa.gov/tax/locgov/Statistical_Calculation_Definitions.pdf
- [5] <http://www.tax.ny.gov/research/property/reports/cod/2010mvs/reporttext.htm>
- [6] <http://www.state.sd.us/drr2/publications/assess1199.pdf>
- [7] Puri, Madan L.; Sen, Pranab K.; *Nonparametric Methods in Multivariate Analysis*, John Wiley & Sons, New York, NY, 1971. (Reprinted by Krieger Publishing)
- [8] <http://www.universityofcalifornia.edu/senate/inmemoriam/georgewbrown.htm>
- [9] Laplace PS de (1818) *Deuxième supplément à la Théorie Analytique des Probabilités*, Paris, Courcier
- [10] Keynes, J.M. (1921) *A Treatise on Probability*. Pt II Ch XVII §5 (p 201) (2006 reprint, Cosimo Classics, ISBN 9781596055308 : multiple other reprints)
- [11] Galton F (1881) "Report of the Anthropometric Committee" pp 245-260. *Report of the 51st Meeting of the British Association for the Advancement of Science* (<http://www.biodiversitylibrary.org/item/94448>)

External links

- Hazewinkel, Michiel, ed. (2001), "Median (in statistics)" (<http://www.encyclopediaofmath.org/index.php?title=p/m063310>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- A Guide to Understanding & Calculating the Median (<http://stats4students.com/measures-of-central-tendency-2.php>)
- Median as a weighted arithmetic mean of all Sample Observations (<http://www.accessecon.com/pubs/EB/2004/Volume3/EB-04C10011A.pdf>)
- On-line calculator (<http://www.poorcity.richcity.org/cgi-bin/inequality.cgi>)
- Calculating the median (<http://www.statcan.gc.ca/edu/power-pouvoir/ch11/median-médiane/5214872-eng.htm>)
- A problem involving the mean, the median, and the mode. (http://mathschallenge.net/index.php?section=problems&show=true&titleid=average_problem)
- Weisstein, Eric W., "Statistical Median" (<http://mathworld.wolfram.com/StatisticalMedian.html>), *MathWorld*.
- Python script (<http://www.poorcity.richcity.org/oei/#GiniHooverTheil>) for Median computations and income inequality metrics

This article incorporates material from Median of a distribution on PlanetMath, which is licensed under the Creative Commons Attribution/Share-Alike License.

Mode (statistics)

The **mode** is the value that appears most often in a set of data. The mode of a discrete probability distribution is the value x at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled. The mode of a continuous probability distribution is the value x at which its probability density function has its maximum value, so, informally speaking, the mode is at the peak.

Like the statistical mean and median, the mode is a way of expressing, in a single number, important information about a random variable or a population. The numerical value of the mode is the same as that of the mean and median in a normal distribution, and it may be very different in highly skewed distributions.

The mode is not necessarily unique, since the probability mass function or probability density function may take the same maximum value at several points x_1, x_2 , etc. The most extreme case occurs in uniform distributions, where all values occur equally frequently.

The above definition tells us that only *global maxima* are modes. Slightly confusingly, when a probability density function has multiple local maxima it is common to refer to all of the local maxima as modes of the distribution. Such a continuous distribution is called multimodal (as opposed to unimodal).

In symmetric unimodal distributions, such as the normal (or Gaussian) distribution (the distribution whose density function, when graphed, gives the famous "bell curve"), the mean (if defined), median and mode all coincide. For samples, if it is known that they are drawn from a symmetric distribution, the sample mean can be used as an estimate of the population mode.

Mode of a sample

The mode of a sample is the element that occurs most often in the collection. For example, the mode of the sample [1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17] is 6. Given the list of data [1, 1, 2, 4, 4] the mode is not unique - the dataset may be said to be bimodal, while a set with more than two modes may be described as multimodal.

For a sample from a continuous distribution, such as [0.935..., 1.211..., 2.430..., 3.668..., 3.874...], the concept is unusable in its raw form, since no two values will be exactly the same, so each value will occur precisely once. In order to estimate the mode, the usual practice is to discretize the data by assigning frequency values to intervals of equal distance, as for making a histogram, effectively replacing the values by the midpoints of the intervals they are assigned to. The mode is then the value where the histogram reaches its peak. For small or middle-sized samples the outcome of this procedure is sensitive to the choice of interval width if chosen too narrow or too wide; typically one should have a sizable fraction of the data concentrated in a relatively small number of intervals (5 to 10), while the fraction of the data falling outside these intervals is also sizable. An alternate approach is kernel density estimation, which essentially blurs point samples to produce a continuous estimate of the probability density function which can provide an estimate of the mode.

The following MATLAB (or Octave) code example computes the mode of a sample:

```
X = sort(x);
indices = find(diff([X; realmax]) > 0); % indices where repeated
values change
[modeL,i] = max(diff([0; indices])); % longest persistence length
of repeated values
mode = X(indices(i));
```

The algorithm requires as a first step to sort the sample in ascending order. It then computes the discrete derivative of the sorted list, and finds the indices where this derivative is positive. Next it computes the discrete derivative of this set of indices, locating the maximum of this derivative of indices, and finally evaluates the sorted sample at the point where that maximum occurs, which corresponds to the last member of the stretch of repeated values.

Comparison of mean, median and mode

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$(1+2+2+3+4+7+9) / 7$	4
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, 3 , 4, 7, 9	3
Mode	Most frequent value in a data set	1, 2 , 2, 3, 4, 7, 9	2

Use

Unlike mean and median, the concept of mode also makes sense for "nominal data" (i.e., not consisting of numerical values in the case of mean, or even of ordered values in the case of median). For example, taking a sample of Korean family names, one might find that "Kim" occurs more often than any other name. Then "Kim" would be the mode of the sample. In any voting system where a plurality determines victory, a single modal value determines the victor, while a multi-modal outcome would require some tie-breaking procedure to take place.

Unlike median, the concept of mean makes sense for any random variable assuming values from a vector space, including the real numbers (a one-dimensional vector space) and the integers (which can be considered embedded in the reals). For example, a distribution of points in the plane will typically have a mean and a mode, but the concept of median does not apply. The median makes sense when there is a linear order on the possible values. Generalizations of the concept of median to higher-dimensional spaces are the geometric median and the centerpoint.

Uniqueness and definedness

For the remainder, the assumption is that we have (a sample of) a real-valued random variable.

For some probability distributions, the expected value may be infinite or undefined, but if defined, it is unique. The mean of a (finite) sample is always defined. The median is the value such that the fractions not exceeding it and not falling below it are both at least 1/2. It is not necessarily unique, but never infinite or totally undefined. For a data sample it is the "halfway" value when the list of values is ordered in increasing value, where usually for a list of even length the numerical average is taken of the two values closest to "halfway". Finally, as said before, the mode is not necessarily unique. Certain pathological distributions (for example, the Cantor distribution) have no defined mode at all.^[citation needed] For a finite data sample, the mode is one (or more) of the values in the sample.

Properties

Assuming definedness, and for simplicity uniqueness, the following are some of the most interesting properties.

- All three measures have the following property: If the random variable (or each value from the sample) is subjected to the linear or affine transformation which replaces X by $aX+b$, so are the mean, median and mode.
- However, if there is an arbitrary monotonic transformation, only the median follows; for example, if X is replaced by $\exp(X)$, the median changes from m to $\exp(m)$ but the mean and mode won't.^[citation needed]
- Except for extremely small samples, the mode is insensitive to "outliers" (such as occasional, rare, false experimental readings). The median is also very robust in the presence of outliers, while the mean is rather sensitive.

- In continuous unimodal distributions the median lies, as a rule of thumb, between the mean and the mode, about one third of the way going from mean to mode. In a formula, $\text{median} \approx (2 \times \text{mean} + \text{mode})/3$. This rule, due to Karl Pearson, often applies to slightly non-symmetric distributions that resemble a normal distribution, but it is not always true and in general the three statistics can appear in any order.
- For unimodal distributions, the mode is within $\sqrt{3}$ standard deviations of the mean, and the root mean square deviation about the mode is between the standard deviation and twice the standard deviation.

Example for a skewed distribution

An example of a skewed distribution is personal wealth: Few people are very rich, but among those some are extremely rich. However, many are rather poor.

A well-known class of distributions that can be arbitrarily skewed is given by the log-normal distribution. It is obtained by transforming a random variable X having a normal distribution into random variable $Y = e^X$. Then the logarithm of random variable Y is normally distributed, hence the name.

Taking the mean μ of X to be 0, the median of Y will be 1, independent of the standard deviation σ of X . This is so because X has a symmetric distribution, so its median is also 0. The transformation from X to Y is monotonic, and so we find the median $e^0 = 1$ for Y .

When X has standard deviation $\sigma = 0.25$, the distribution of Y is weakly skewed. Using formulas for the log-normal distribution, we find:

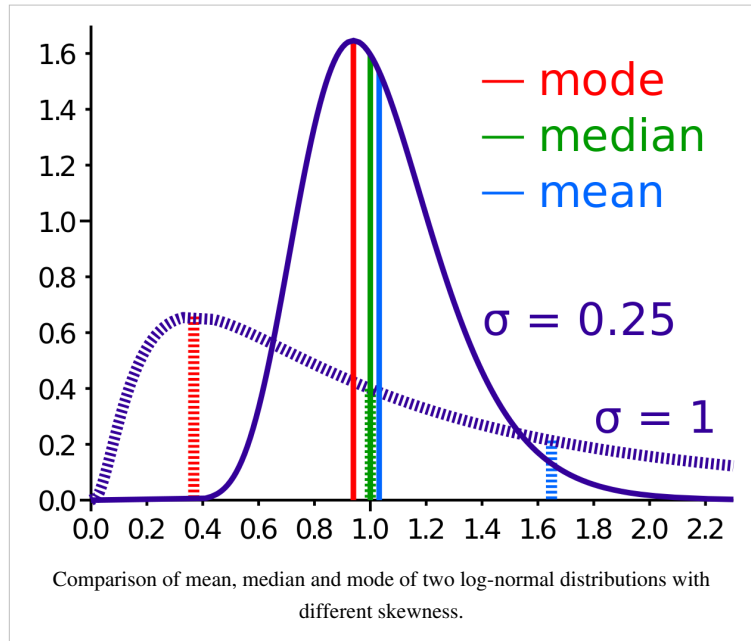
$$\begin{aligned} \text{mean} &= e^{\mu+\sigma^2/2} = e^{0+0.25^2/2} \approx 1.032 \\ \text{mode} &= e^{\mu-\sigma^2} = e^{0-0.25^2} \approx 0.939 \\ \text{median} &= e^{\mu} = e^0 = 1 \end{aligned}$$

Indeed, the median is about one third on the way from mean to mode.

When X has a larger standard deviation, $\sigma = 1$, the distribution of Y is strongly skewed. Now

$$\begin{aligned} \text{mean} &= e^{\mu+\sigma^2/2} = e^{0+1^2/2} \approx 1.649 \\ \text{mode} &= e^{\mu-\sigma^2} = e^{0-1^2} \approx 0.368 \\ \text{median} &= e^{\mu} = e^0 = 1 \end{aligned}$$

Here, Pearson's rule of thumb fails.



van Zwet condition

Van Zwet derived an inequality which provides sufficient conditions for this inequality to hold.^[1] The inequality

$$\text{Mode} \leq \text{Median} \leq \text{Mean}$$

holds if

$$F(\text{Median} - x) + F(\text{Median} + x) \geq 1$$

for all x where $F()$ is the cumulative distribution function of the distribution.

Unimodal distributions

The difference between the mean and the mode in a unimodal continuous distribution is bounded by the standard deviation multiplied by the square root of 3.^[2] In symbols

$$\frac{|\text{mean} - \text{mode}|}{\text{standard deviation}} \leq \sqrt{3}$$

where $||$ is the absolute value. Incidentally this formula is also the Pearson mode or first skewness coefficient.

The difference between the mode and the median has the same bound. In symbols

$$\frac{|\text{median} - \text{mode}|}{\text{standard deviation}} \leq \sqrt{3}$$

Confidence interval for the mode with a single data point

It is a common but false belief that from a single observation x we can not gain information about the variability in the population and that consequently that finite length confidence intervals for mean and/or variance are impossible even in principle.

It is possible for an unknown unimodal distribution to estimate a confidence interval for the mode with a sample size of 1. This was first shown by Abbot and Rosenblatt and extended by Blachman and Machol. This confidence interval can be sharpened if the distribution can be assumed to be symmetrical. It is further possible to sharpen this interval if the distribution is normally distributed.

Let the confidence interval be $1 - \alpha$. Then the confidence intervals for the general, symmetric and normally distributed variates respectively are

$$X \pm \left(\frac{2}{\alpha} - 1\right) |X - \theta|$$

$$X \pm \left(\frac{1}{\alpha} - 1\right) |X - \theta|$$

$$X \pm \left(\frac{0.484}{\alpha} - 1\right) |X - \theta|$$

where X is the variate, θ is the mode and $||$ is the absolute value.

These estimates are conservative. The confidence intervals for the mode at the 90% level given by these estimators are $X \pm 19 |X - \theta|$, $X \pm 9 |X - \theta|$ and $X \pm 5.84 |X - \theta|$ for the general, symmetric and normally distributed variates respectively. The 95% confidence interval for a normally distributed variate is given by $X \pm 10.7 |X - \theta|$. It may be worth noting that the mean and the mode coincide if the variates are normally distributed.

The 95% bound for a normally distributed variate has been improved and is now known to be $X \pm 9.68 |X - \theta|$. The bound for a 99% confidence interval is $X \pm 48.39 |X - \theta|$

Note

Machol has shown that that given a known density symmetrical about 0 that given a single sample value (x) that the 90% confidence intervals of population mean are^[3]

$$x \pm 5|x - \nu|$$

where ν is the population median.

If the precise form of the distribution is not known but it is known to be symmetrical about zero then we have

$$P(X - k|X - a| \leq \mu \leq X + k|X - a|) \geq 1 - \frac{1}{1 + k}$$

where X is the variate, μ is the population mean and a and k are arbitrary real numbers.

It is also possible to estimate a confidence interval for the standard deviation from a single observation if the distribution is symmetrical about 0. For a normal distribution the with an unknown variance and a single data point (X) the 90%, 95% and 99% confidence intervals for the standard deviation are [0, 8|X|], [0, 17|X|] and [0, 70|X|]. These intervals may be shorted if the mean is known to be bounded by a multiple of the standard deviation.

If the distribution is known to be normal then it is possible to estimate a confidence interval for both the mean and variance from a simple value. The 90% confidence intervals are

$$\begin{aligned} X - 23.3|X| &\leq \mu \leq X + 23.3|X| \\ \sigma &\leq 10|X| \end{aligned}$$

The confidence intervals can be estimated for any chosen range.

This method is not limited to the normal distribution but can be used with any known distribution.

Statistical tests

These estimators have been used to create hypothesis tests for simple samples from normal or symmetrical unimodal distributions. Let the distribution have an assumed mean (μ_0). The null hypothesis is that the assumed mean of the distribution lies within the confidence interval of the sample mean (m). The null hypothesis is accepted if

$$\mu_0 < \frac{x + m}{2} \pm k|x - m|$$

where x is the value of the sample and k is a constant. The null hypothesis is rejected if

$$\mu_0 > \frac{x + m}{2} \pm k|x - m|$$

The value of k depends on the choice of confidence interval and the nature of the assumed distribution.

If the distribution is assumed or is known to be normal then the values of k for the 50%, 66.6%, 75%, 80%, 90%, 95% and 99% confidence intervals are 0.50, 1.26, 1.80, 2.31, 4.79, 9.66 and 48.39 respectively.

If the distribution is assumed or known to be unimodal and symmetrical but not normal then the values of k for the 50%, 66.6%, 75%, 80%, 90%, 95% and 99% confidence intervals are 0.50, 1.87, 2.91, 3.94, 8.97, 18.99, 99.00 respectively.

To see how this test works we assume or know *a priori* that the population from which the sample is drawn has a mean of μ_0 and that the population has a symmetrical unimodal distribution - a class that includes the normal distribution. We wish to know if the mean estimated from the sample is representative of the population at a pre chosen level of confidence.

Assume that the distribution is normal and let the confidence interval be 95%. Then $k = 9.66$.

Assuming that the sample is representative of the population, the sample mean (m) will then lie within the range determined from the formula:

$$\mu_0 < \frac{x + m}{2} \pm 9.66|x - m|$$

If subsequent sampling shows that the sample mean lies outside these parameters the sample mean is to be considered to differ significantly from the population mean.

History

The term mode originates with Karl Pearson in 1895.^[4]

References

- [1] van Zwet WR (1979) "Mean, median, mode II", *Statistica Neerlandica*, 33 (1) 1–5
- [2] <http://www.se16.info/hgb/cheb2.htm#3unimodalequalities>
- [3] Machol R (1964) IEEE Trans Info Theor
- [4] Pearson, Karl (1895). "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material", *Philosophical Transactions of the Royal Society of London, Ser. A*, 186, 343-414

External links

- Hazewinkel, Michiel, ed. (2001), "Mode" (<http://www.encyclopediaofmath.org/index.php?title=p/m064340>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- A Guide to Understanding & Calculating the Mode (http://www.stats4students.com/Essentials/Measures-Central-Tendency/Overview_2.php)
- Weisstein, Eric W., " Mode (<http://mathworld.wolfram.com/Mode.html>)", *MathWorld*.
- Mean, Median and Mode short beginner video from Khan Academy (<http://www.khanacademy.org/math/statistics/v/mean-median-and-mode>)

Variance

In probability theory and statistics, **variance** measures how far a set of numbers is spread out. (A variance of zero indicates that all the values are identical.) A non-zero variance is always positive: A small variance indicates that the data points tend to be very close to the mean (expected value) and hence to each other, while a high variance indicates that the data points are very spread out from the mean and from each other.

The square root of variance is called the standard deviation.

The variance is one of several descriptors of a probability distribution. In particular, the variance is one of the moments of a distribution. In that context, it forms part of a systematic approach to distinguishing between probability distributions. While other such approaches have been developed, those based on moments are advantageous in terms of mathematical and computational simplicity.

The variance is a parameter that describes, in part, either the actual probability distribution of an observed population of numbers, or the theoretical probability distribution of a sample (a not-fully-observed population) of numbers. In the latter case, a sample of data from such a distribution can be used to construct an estimate of its variance: in the simplest cases this estimate can be the sample variance.

Definition

The variance of a random variable X is its second central moment, the expected value of the squared deviation from the mean $\mu = E[X]$:

$$\text{Var}(X) = E[(X - \mu)^2].$$

This definition encompasses random variables that are discrete, continuous, neither, or mixed. The variance can also be thought of as the covariance of a random variable with itself:

$$\text{Var}(X) = \text{Cov}(X, X).$$

The variance is also equivalent to the second cumulant of the probability distribution for X . The variance is typically designated as $\text{Var}(X)$, σ_X^2 , or simply σ^2 (pronounced "sigma squared"). The expression for the variance can be

expanded:

$$\begin{aligned}\text{Var}(X) &= \mathbb{E} [X^2 - 2XE[X] + (E[X])^2] \\ &= \mathbb{E} [X^2] - 2E[X]E[X] + (E[X])^2 \\ &= \mathbb{E} [X^2] - (E[X])^2\end{aligned}$$

A mnemonic for the above expression is "mean of square minus square of mean".

Continuous random variable

If the random variable X is continuous with probability density function $f(x)$, then the variance is given by

$$\text{Var}(X) = \sigma^2 = \int (x - \mu)^2 f(x) dx = \int x^2 f(x) dx - \mu^2$$

where μ is the expected value,

$$\mu = \int x f(x) dx$$

and where the integrals are definite integrals taken for x ranging over the range of X .

If a continuous distribution does not have an expected value, as is the case for the Cauchy distribution, it does not have a variance either. Many other distributions for which the expected value does exist also do not have a finite variance because the integral in the variance definition diverges. An example is a Pareto distribution whose index k satisfies $1 < k \leq 2$.

Discrete random variable

If the random variable X is discrete with probability mass function $x_1 \mapsto p_1, \dots, x_n \mapsto p_n$, then

$$\text{Var}(X) = \sum_{i=1}^n p_i \cdot (x_i - \mu)^2 = \sum_{i=1}^n (p_i \cdot x_i^2) - \mu^2$$

where μ is the expected value, i.e.

$$\mu = \sum_{i=1}^n p_i \cdot x_i.$$

(When such a discrete weighted variance is specified by weights whose sum is not 1, then one divides by the sum of the weights.)

The variance of a set of n equally likely values can be written as

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2.$$

The variance of a set of n equally likely values can be equivalently expressed, without directly referring to the mean, in terms of squared deviations of all points from each other:

$$\text{Var}(X) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \frac{1}{2} (x_i - x_j)^2.$$

Examples

Normal distribution

The normal distribution with parameters μ and σ is a continuous distribution whose probability density function is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

It has mean μ and variance equal to:

$$\text{Var}(X) = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2.$$

The role of the normal distribution in the central limit theorem is in part responsible for the prevalence of the variance in probability and statistics.

Exponential distribution

The exponential distribution with parameter λ is a continuous distribution whose support is the semi-infinite interval $[0, \infty)$. Its probability density function is given by:

$$f(x) = \lambda e^{-\lambda x},$$

and it has expected value $\mu = \lambda^{-1}$. The variance is equal to:

$$\text{Var}(X) = \int_0^{\infty} (x - \lambda^{-1})^2 \lambda e^{-\lambda x} dx = \lambda^{-2}.$$

So for an exponentially distributed random variable $\sigma^2 = \mu^2$.

Poisson distribution

The Poisson distribution with parameter λ is a discrete distribution for $k = 0, 1, 2, \dots$. Its probability mass function is given by:

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

and it has expected value $\mu = \lambda$. The variance is equal to:

$$\text{Var}(X) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} (k - \lambda)^2 = \lambda,$$

So for a Poisson-distributed random variable $\sigma^2 = \mu$.

Binomial distribution

The binomial distribution with parameters n and p is a discrete distribution for $k = 0, 1, 2, \dots, n$. Its probability mass function is given by:

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

and it has expected value $\mu = np$. The variance is equal to:

$$\text{Var}(X) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} (k - np)^2 = np(1-p),$$

Coin toss

The binomial distribution with $p = 0.5$ describes the probability of getting k heads in n tosses. Thus the expected value of the number of heads is $\frac{n}{2}$, and the variance is $\frac{n}{4}$.

Fair die

A six-sided fair die can be modelled with a discrete random variable with outcomes 1 through 6, each with equal probability $\frac{1}{6}$. The expected value is $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$. Therefore the variance can be computed to be:

$$\begin{aligned} \sum_{i=1}^6 \frac{1}{6}(i - 3.5)^2 &= \frac{1}{6} \sum_{i=1}^6 (i - 3.5)^2 = \frac{1}{6} ((-2.5)^2 + (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 + 2.5^2) \\ &= \frac{1}{6} \cdot 17.50 = \frac{35}{12} \approx 2.92. \end{aligned}$$

The general formula for the variance of the outcome X of a die of n sides is:

$$\begin{aligned} \sigma^2 = E(X^2) - (E(X))^2 &= \frac{1}{n} \sum_{i=1}^n i^2 - \left(\frac{1}{n} \sum_{i=1}^n i \right)^2 \\ &= \frac{1}{6}(n + 1)(2n + 1) - \frac{1}{4}(n + 1)^2 \\ &= \frac{n^2 - 1}{12}. \end{aligned}$$

Properties

Basic properties

Variance is non-negative because the squares are positive or zero.

$$\text{Var}(X) \geq 0.$$

The variance of a constant random variable is zero, and if the variance of a variable in a data set is 0, then all the entries have the same value.

$$P(X = a) = 1 \Leftrightarrow \text{Var}(X) = 0.$$

Variance is invariant with respect to changes in a location parameter. That is, if a constant is added to all values of the variable, the variance is unchanged.

$$\text{Var}(X + a) = \text{Var}(X).$$

If all values are scaled by a constant, the variance is scaled by the square of that constant.

$$\text{Var}(aX) = a^2 \text{Var}(X).$$

The variance of a sum of two random variables is given by:

$$\begin{aligned} \text{Var}(aX + bY) &= a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y), \\ \text{Var}(X - Y) &= \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y), \end{aligned}$$

where $\text{Cov}(\cdot, \cdot)$ is the covariance. In general we have for the sum of N random variables $\{X_1, \dots, X_N\}$:

$$\text{Var} \left(\sum_{i=1}^N X_i \right) = \sum_{i,j=1}^N \text{Cov}(X_i, X_j) = \sum_{i=1}^N \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j).$$

These results lead to the variance of a linear combination as:

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^N a_i X_i\right) &= \sum_{i,j=1}^N a_i a_j \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^N a_i^2 \text{Var}(X_i) + \sum_{i \neq j} a_i a_j \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^N a_i^2 \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq N} a_i a_j \text{Cov}(X_i, X_j).
\end{aligned}$$

If the random variables X_1, \dots, X_N are such that

$$\text{Cov}(X_i, X_j) = 0, \quad \forall (i \neq j),$$

they are said to be uncorrelated. It follows immediately from the expression given earlier that if the random variables X_1, \dots, X_N are uncorrelated, then the variance of their sum is equal to the sum of their variances, or, expressed symbolically:

$$\text{Var}\left(\sum_{i=1}^N X_i\right) = \sum_{i=1}^N \text{Var}(X_i).$$

Since independent random variables are always uncorrelated, the equation above holds in particular when the random variables X_1, \dots, X_n are independent. Thus independence is sufficient but not necessary for the variance of the sum to equal the sum of the variances.

Sum of uncorrelated variables (Bienaymé formula)

One reason for the use of the variance in preference to other measures of dispersion is that the variance of the sum (or the difference) of uncorrelated random variables is the sum of their variances:

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i).$$

This statement is called the Bienaymé formula^[1] and was discovered in 1853.^[citation needed] It is often made with the stronger condition that the variables are independent, but uncorrelatedness suffices. So if all the variables have the same variance σ^2 , then, since division by n is a linear transformation, this formula immediately implies that the variance of their mean is

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}.$$

That is, the variance of the mean decreases when n increases. This formula for the variance of the mean is used in the definition of the standard error of the sample mean, which is used in the central limit theorem.

Product of independent variables

If two variables X and Y are independent, the variance of their product is given by^{[2][3]}

$$\begin{aligned} \text{Var}(XY) &= [E(X)]^2\text{Var}(Y) + [E(Y)]^2\text{Var}(X) + \text{Var}(X)\text{Var}(Y) \\ &= E(X^2)E(Y^2) - [E(X)]^2[E(Y)]^2. \end{aligned}$$

Sum of correlated variables

In general, if the variables are correlated, then the variance of their sum is the sum of their covariances:

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j).$$

(Note: The second equality comes from the fact that $\text{Cov}(X_i, X_i) = \text{Var}(X_i)$.)

Here Cov is the covariance, which is zero for independent random variables (if it exists). The formula states that the variance of a sum is equal to the sum of all elements in the covariance matrix of the components. This formula is used in the theory of Cronbach's alpha in classical test theory.

So if the variables have equal variance σ^2 and the average correlation of distinct variables is ρ , then the variance of their mean is

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} + \frac{n-1}{n} \rho \sigma^2.$$

This implies that the variance of the mean increases with the average of the correlations. Moreover, if the variables have unit variance, for example if they are standardized, then this simplifies to

$$\text{Var}(\bar{X}) = \frac{1}{n} + \frac{n-1}{n} \rho.$$

This formula is used in the Spearman–Brown prediction formula of classical test theory. This converges to ρ if n goes to infinity, provided that the average correlation remains constant or converges too. So for the variance of the mean of standardized variables with equal correlations or converging average correlation we have

$$\lim_{n \rightarrow \infty} \text{Var}(\bar{X}) = \rho.$$

Therefore, the variance of the mean of a large number of standardized variables is approximately equal to their average correlation. This makes clear that the sample mean of correlated variables does generally not converge to the population mean, even though the Law of large numbers states that the sample mean will converge for independent variables.

Weighted sum of variables

The scaling property and the Bienaymé formula, along with this property from the covariance page: $\text{Cov}(aX, bY) = ab \text{Cov}(X, Y)$ jointly imply that

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab \text{Cov}(X, Y).$$

This implies that in a weighted sum of variables, the variable with the largest weight will have a disproportionately large weight in the variance of the total. For example, if X and Y are uncorrelated and the weight of X is two times the weight of Y , then the weight of the variance of X will be four times the weight of the variance of Y .

The expression above can be extended to a weighted sum of multiple variables:

$$\text{Var}\left(\sum_i^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j \text{Cov}(X_i, X_j)$$

Decomposition

The general formula for variance decomposition or the law of total variance is: If X and Y are two random variables, and the variance of X exists, then

$$\text{Var}(X) = \text{Var}(E(X|Y)) + E(\text{Var}(X|Y)).$$

Here, $E(X|Y)$ is the conditional expectation of X given Y , and $\text{Var}(X|Y)$ is the conditional variance of X given Y . (A more intuitive explanation is that given a particular value of Y , then X follows a distribution with mean $E(X|Y)$ and variance $\text{Var}(X|Y)$. The above formula tells how to find $\text{Var}(X)$ based on the distributions of these two quantities when Y is allowed to vary.) This formula is often applied in analysis of variance, where the corresponding formula is

$$MS_{\text{total}} = MS_{\text{between}} + MS_{\text{within}};$$

here MS refers to the Mean of the Squares. It is also used in linear regression analysis, where the corresponding formula is

$$MS_{\text{total}} = MS_{\text{regression}} + MS_{\text{residual}}.$$

This can also be derived from the additivity of variances, since the total (observed) score is the sum of the predicted score and the error score, where the latter two are uncorrelated.

Similar decompositions are possible for the sum of squared deviations (sum of squares, SS):

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}};$$

$$SS_{\text{total}} = SS_{\text{regression}} + SS_{\text{residual}}.$$

Formulae for the variance

A formula often used for deriving the variance of a theoretical distribution is as follows:

$$\text{Var}(X) = E(X^2) - (E(X))^2.$$

This will be useful when it is possible to derive formulae for the expected value and for the expected value of the square.

This formula is also sometimes used in connection with the sample variance. While useful for hand calculations, it is not advised for computer calculations as it suffers from catastrophic cancellation if the two components of the equation are similar in magnitude and floating point arithmetic is used. This is discussed in the article Algorithms for calculating variance.

Calculation from the CDF

The population variance for a non-negative random variable can be expressed in terms of the cumulative distribution function F using

$$2 \int_0^{\infty} uH(u) du - \left(\int_0^{\infty} H(u) du \right)^2.$$

where $H(u) = 1 - F(u)$ is the right tail function. This expression can be used to calculate the variance in situations where the CDF, but not the density, can be conveniently expressed.

Characteristic property

The second moment of a random variable attains the minimum value when taken around the first moment (i.e., mean) of the random variable, i.e. $\operatorname{argmin}_m \mathbb{E}((X - m)^2) = \mathbb{E}(X)$. Conversely, if a continuous function φ satisfies $\operatorname{argmin}_m \mathbb{E}(\varphi(X - m)) = \mathbb{E}(X)$ for all random variables X , then it is necessarily of the form $\varphi(x) = ax^2 + b$, where $a > 0$. This also holds in the multidimensional case.

Matrix notation for the variance of a linear combination

Let's define X as a column vector of n random variables X_1, \dots, X_n , and c as a column vector of N scalars c_1, \dots, c_n . Therefore $c^T X$ is a linear combination of these random variables, where c^T denotes the transpose of vector c . Let also be Σ the variance-covariance matrix of the vector X . The variance of $c^T X$ is given by:

$$\operatorname{Var}(c^T X) = c^T \Sigma c.$$

Units of measurement

Unlike expected absolute deviation, the variance of a variable has units that are the square of the units of the variable itself. For example, a variable measured in inches will have a variance measured in square inches. For this reason, describing data sets via their standard deviation or root mean square deviation is often preferred over using the variance. In the dice example the standard deviation is $\sqrt{2.9} \approx 1.7$, slightly larger than the expected absolute deviation of 1.5.

The standard deviation and the expected absolute deviation can both be used as an indicator of the "spread" of a distribution. The standard deviation is more amenable to algebraic manipulation than the expected absolute deviation, and, together with variance and its generalization covariance, is used frequently in theoretical statistics; however the expected absolute deviation tends to be more robust as it is less sensitive to outliers arising from measurement anomalies or an unduly heavy-tailed distribution.

Approximating the variance of a function

The delta method uses second-order Taylor expansions to approximate the variance of a function of one or more random variables: see Taylor expansions for the moments of functions of random variables. For example, the approximate variance of a function of one variable is given by

$$\operatorname{Var}[f(X)] \approx (f'(E[X]))^2 \operatorname{Var}[X]$$

provided that f is twice differentiable and that the mean and variance of X are finite.

Population variance and sample variance

Real-world distributions such as the distribution of yesterday's rain throughout the day are typically not fully known, unlike the behavior of perfect dice or an ideal distribution such as the normal distribution, because it is impractical to account for every raindrop. Instead one estimates the mean and variance of the whole distribution as the computed mean and variance of a sample of n observations drawn suitably randomly from the whole sample space, in this example the set of all measurements of yesterday's rainfall in all available rain gauges.

This method of estimation is close to optimal, with the caveat that it underestimates the variance by a factor of $(n - 1) / n$. (For example, when $n = 1$ the variance of a single observation is obviously zero regardless of the true variance). This gives a bias which should be corrected for when n is small by multiplying by $n / (n - 1)$. If the mean is determined in some other way than from the same samples used to estimate the variance then this bias does not arise and the variance can safely be estimated as that of the samples.

Population variance

In general, the *population variance* of a *finite* population of size N with values x_i is given by

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \mu^2$$

where

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

is the population mean. The population variance therefore is the variance of the underlying probability distribution. In this sense, the concept of population can be extended to continuous random variables with infinite populations.

Sample variance

In many practical situations, the true variance of a population is not known *a priori* and must be computed somehow. When dealing with extremely large populations, it is not possible to count every object in the population, so the computation must be performed on a sample of the population.^[4] Sample variance can also be applied to the estimation of the variance of a continuous distribution from a sample of that distribution.

We take a sample with replacement of n values y_1, \dots, y_n from the population, where $n < N$, and estimate the variance on the basis of this sample.^[5] Directly taking the variance of the sample gives:

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

Here, \bar{y} denotes the sample mean:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Since the y_i are selected randomly, both \bar{y} and σ_y^2 are random variables. Their expected values can be evaluated by summing over the ensemble of all possible samples $\{y_i\}$ from the population. For σ_y^2 this gives:

$$\begin{aligned} E[\sigma_y^2] &= E \left[\frac{1}{n} \sum_{i=1}^n \left(y_i - \frac{1}{n} \sum_{j=1}^n y_j \right)^2 \right] \\ &= \frac{1}{n} \sum_{i=1}^n E \left[y_i^2 - \frac{2}{n} y_i \sum_{j=1}^n y_j + \frac{1}{n^2} \sum_{j=1}^n y_j \sum_{k=1}^n y_k \right] \\ &= \frac{1}{n} \sum_{i=1}^n \left[\frac{n-2}{n} E[y_i^2] - \frac{2}{n} \sum_{j \neq i} E[y_i y_j] + \frac{1}{n^2} \sum_{j=1}^n \sum_{k \neq j} E[y_j y_k] + \frac{1}{n^2} \sum_{j=1}^n E[y_j^2] \right] \\ &= \frac{1}{n} \sum_{i=1}^n \left[\frac{n-2}{n} (\sigma^2 + \mu^2) - \frac{2}{n} (n-1) \mu^2 + \frac{1}{n^2} n(n-1) \mu^2 + \frac{1}{n} (\sigma^2 + \mu^2) \right] \\ &= \frac{n-1}{n} \sigma^2. \end{aligned}$$

Hence σ_y^2 gives an estimate of the population variance that is biased by a factor of $(n-1)/n$. For this reason, σ_y^2 is referred to as the *biased sample variance*. Correcting for this bias yields the *unbiased sample variance*:

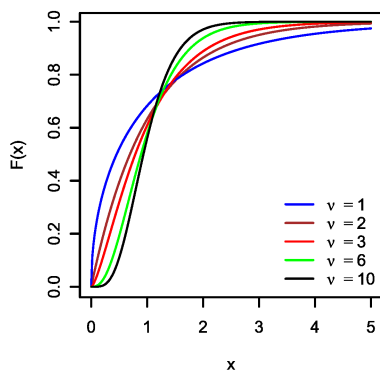
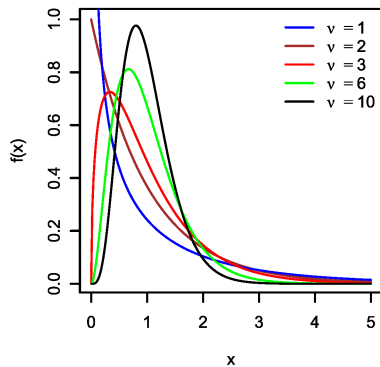
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

Either estimator may be simply referred to as the *sample variance* when the version can be determined by context. The same proof is also applicable for samples taken from a continuous probability distribution.

The use of the term $n - 1$ is called Bessel's correction, and it is also used in sample covariance and the sample standard deviation (the square root of variance). The square root is a concave function and thus introduces negative bias (by Jensen's inequality), which depends on the distribution, and thus the corrected sample standard deviation (using Bessel's correction) is biased. The unbiased estimation of standard deviation is a technically involved problem, though for the normal distribution using the term $n - 1.5$ yields an almost unbiased estimator.

The unbiased sample variance is a U-statistic for the function $f(y_1, y_2) = (y_1 - y_2)^2/2$, meaning that it is obtained by averaging a 2-sample statistic over 2-element subsets of the population.

Distribution of the sample variance



Distribution and cumulative distribution of s^2/σ^2 , for various values of $\nu = n - 1$, when the y_i are independent normally distributed.

Being a function of random variables, the sample variance is itself a random variable, and it is natural to study its distribution. In the case that y_i are independent observations from a normal distribution, Cochran's theorem shows that s^2 follows a scaled chi-squared distribution:^[6]

$$(n - 1) \frac{s^2}{\sigma^2} \sim \chi_{n-1}^2.$$

As a direct consequence, it follows that

$$E(s^2) = E\left(\frac{\sigma^2}{n - 1} \chi_{n-1}^2\right) = \sigma^2,$$

and^[7]

$$\text{Var}[s^2] = \text{Var}\left(\frac{\sigma^2}{n - 1} \chi_{n-1}^2\right) = \frac{\sigma^4}{(n - 1)^2} \text{Var}(\chi_{n-1}^2) = \frac{2\sigma^4}{n - 1}.$$

If the y_i are independent and identically distributed, but not necessarily normally distributed, then^[8]

$$\mathbf{E}[s^2] = \sigma^2, \quad \text{Var}[s^2] = \sigma^4 \left(\frac{2}{n-1} + \frac{\kappa}{n} \right) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right),$$

where κ is the excess kurtosis of the distribution and μ_4 is the fourth moment about the mean.

If the conditions of the law of large numbers hold for the squared observations, s^2 is a consistent estimator of σ^2 .^[citation needed] One can see indeed that the variance of the estimator tends asymptotically to zero.

Samuelson's inequality

Samuelson's inequality is a result that states bounds on the values that individual observations in a sample can take, given that the sample mean and (biased) variance have been calculated.^[9] Values must lie within the limits $\bar{y} \pm \sigma_y(n-1)^{1/2}$.

Relations with the harmonic and arithmetic means

It has been shown^[10] that for a sample $\{y_i\}$ of real numbers,

$$\sigma_y^2 \leq 2y_{\max}(A - H),$$

where y_{\max} is the maximum of the sample, A is the arithmetic mean, H is the harmonic mean of the sample and σ_y^2 is the (biased) variance of the sample.

This bound has been improved on, and it is known that variance is bounded by

$$\sigma_y^2 \leq \frac{y_{\max}(A - H)(y_{\max} - A)}{y_{\max} - H},$$

$$\sigma_y^2 \geq \frac{y_{\min}(A - H)(A - y_{\min})}{H - y_{\min}},$$

where y_{\min} is the minimum of the sample.^[11]

Generalizations

If \mathbf{X} is a vector-valued random variable, with values in \mathbb{R}^n , and thought of as a column vector, then the natural generalization of variance is $\mathbf{E}((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T)$, where $\boldsymbol{\mu} = \mathbf{E}(\mathbf{X})$ and \mathbf{X}^T is the transpose of \mathbf{X} , and so is a row vector. This variance is a positive semi-definite square matrix, commonly referred to as the covariance matrix.

If \mathbf{X} is a complex-valued random variable, with values in \mathbb{C} , then its variance is $\mathbf{E}((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\dagger)$, where \mathbf{X}^\dagger is the conjugate transpose of \mathbf{X} . This variance is also a positive semi-definite square matrix.

Tests of equality of variances

Testing for the equality of two or more variances is difficult. The F test and chi square tests are both adversely affected by non-normality and are not recommended for this purpose.

Several non parametric tests have been proposed: these include the Barton-David-Ansari-Fruend-Siegel-Tukey test, the Capon test, Mood test, the Klotz test and the Sukhatme test. The Sukhatme test applies to two variances and requires that both medians be known and equal to zero. The Mood, Klotz, Capon and Barton-David-Ansari-Fruend-Siegel-Tukey tests also apply to two variances. They allow the median to be unknown but do require that the two medians are equal.

The Lehman test is a parametric test of two variances. Of this test there are several variants known. Other tests of the equality of variances include the Box test, the Box-Anderson test and the Moses test.

Resampling methods, which include the bootstrap and the jackknife, may be used to test the equality of variances.

History

The term *variance* was first introduced by Ronald Fisher in his 1918 paper *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*.^[12]

The great body of available statistics show us that the deviations of a human measurement from its mean follow very closely the Normal Law of Errors, and, therefore, that the variability may be uniformly measured by the standard deviation corresponding to the square root of the mean square error. When there are two independent causes of variability capable of producing in an otherwise uniform population distributions with standard deviations θ_1 and θ_2 , it is found that the distribution, when both causes act together, has a standard deviation $\sqrt{\theta_1^2 + \theta_2^2}$. It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance...

Moment of inertia

The variance of a probability distribution is analogous to the moment of inertia in classical mechanics of a corresponding mass distribution along a line, with respect to rotation about its center of mass.^[citation needed] It is because of this analogy that such things as the variance are called *moments* of probability distributions.^[citation needed] The covariance matrix is related to the moment of inertia tensor for multivariate distributions. The moment of inertia of a cloud of n points with a covariance matrix of Σ is given by^[citation needed]

$$I = n(\mathbf{1}_{3 \times 3} \text{tr}(\Sigma) - \Sigma).$$

This difference between moment of inertia in physics and in statistics is clear for points that are gathered along a line. Suppose many points are close to the x axis and distributed along it. The covariance matrix might look like

$$\Sigma = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}.$$

That is, there is the most variance in the x direction. However, physicists would consider this to have a low moment *about* the x axis so the moment-of-inertia tensor is

$$I = n \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 10.1 & 0 \\ 0 & 0 & 10.1 \end{bmatrix}.$$

Notes

- [1] Loeve, M. (1977) "Probability Theory", *Graduate Texts in Mathematics*, Volume 45, 4th edition, Springer-Verlag, p. 12.
- [2] Goodman, Leo A., "On the exact variance of products," *Journal of the American Statistical Association*, December 1960, 708–713.
- [3] Goodman, Leo A., "The variance of the product of K random variables," *Journal of the American Statistical Association*, March 1962, 54ff.
- [4] Navidi, William (2006) *Statistics for Engineers and Scientists*, McGraw-Hill, pg 14.
- [5] Montgomery, D. C. and Runger, G. C. (1994) *Applied statistics and probability for engineers*, page 201. John Wiley & Sons New York
- [6] Knight K. (2000), *Mathematical Statistics*, Chapman and Hall, New York. (proposition 2.11)
- [7] Casella and Berger (2002) *Statistical Inference*, Example 7.3.3, p. 331
- [8] Neter, Wasserman, and Kutner (1990) *Applied Linear Statistical Models*, 3rd edition, pp. 622-623
- [9] Samuelson, Paul (1968) "How Deviant Can You Be?", *Journal of the American Statistical Association*, 63, number 324 (December, 1968), pp. 1522–1525
- [10] A. McD. Mercer. Bounds for A-G, A-H, G-H, and a family of inequalities of Ky Fan's type, using a general method. *J. Math. Anal. Appl.* 243, 163–173 (2000)
- [11] R. Sharma. Some more inequalities for arithmetic mean, harmonic mean and variance. *J. Math. Inequalities*, 2(1), 109–114 (2008).
- [12] Ronald Fisher (1918) The correlation between relatives on the supposition of Mendelian Inheritance (<http://digital.library.adelaide.edu.au/dspace/bitstream/2440/15097/1/9.pdf>)

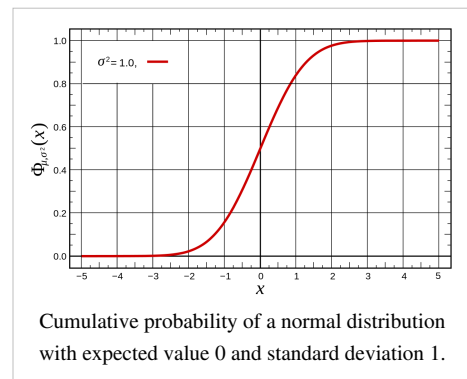
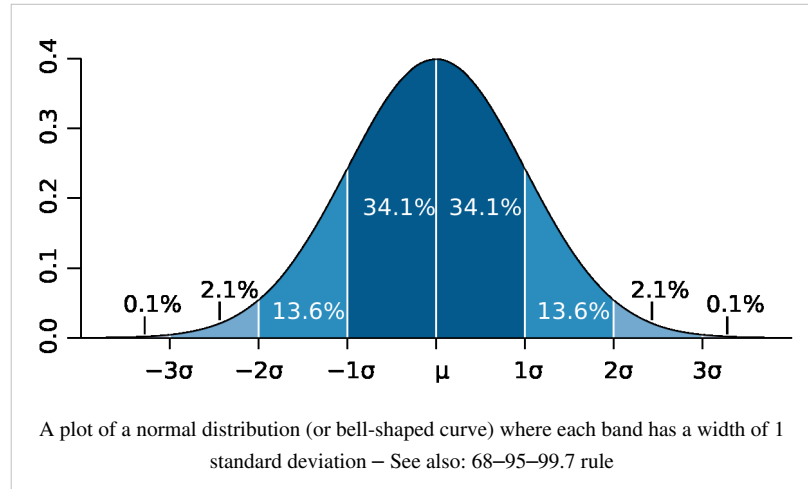
Standard deviation

In statistics and probability theory, the **standard deviation** (represented by the Greek letter sigma, σ) shows how much variation or dispersion from the average exists. A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value); a high standard deviation indicates that the data points are spread out over a large range of values.

The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler though in practice less robust than the average absolute deviation. A useful property of the standard deviation is that, unlike the variance, it is expressed in the same units as the data. Note, however, that for measurements with percentage as the unit, the standard deviation will have percentage points as the unit.

In addition to expressing the variability of a population, the standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times. The reported margin of error is typically about twice the standard deviation – the half-width of a 95 percent confidence interval. In science, researchers commonly report the standard deviation of experimental data, and only effects that fall much farther than one standard deviation away from what would have been expected are considered statistically significant – normal random error or variation in the measurements is in this way distinguished from causal variation. The standard deviation is also important in finance, where the standard deviation on the rate of return on an investment is a measure of the volatility of the investment.

When only a sample of data from a population is available, the term **standard deviation of the sample** or **sample standard deviation** can refer to either the above-mentioned quantity as applied to those data or to a modified quantity that is a better estimate of the **population standard deviation** (the standard deviation of the entire population).



Basic examples

For a finite set of numbers, the standard deviation is found by taking the square root of the average of the squared differences of the values from their average value. For example, consider a **population** consisting of the following eight values:

$$2, 4, 4, 4, 5, 5, 7, 9.$$

These eight data points have the mean (average) of 5:

$$\frac{2 + 4 + 4 + 4 + 5 + 5 + 7 + 9}{8} = 5.$$

First, calculate the difference of each data point from the mean, and square the result of each:

$$\begin{aligned} (2 - 5)^2 &= (-3)^2 = 9 & (5 - 5)^2 &= 0^2 = 0 \\ (4 - 5)^2 &= (-1)^2 = 1 & (5 - 5)^2 &= 0^2 = 0 \\ (4 - 5)^2 &= (-1)^2 = 1 & (7 - 5)^2 &= 2^2 = 4 \\ (4 - 5)^2 &= (-1)^2 = 1 & (9 - 5)^2 &= 4^2 = 16. \end{aligned}$$

Next, calculate the mean of these values, and take the square root:

$$\sqrt{\frac{9 + 1 + 1 + 1 + 0 + 0 + 4 + 16}{8}} = 2.$$

This quantity is the *population* standard deviation, and is equal to the square root of the variance. This formula is valid only if the eight values we began with form the complete population. If the values instead were a random sample drawn from some larger parent population, then we would have divided by 7 (which is $n-1$) instead of 8 (which is n) in the denominator of the last formula, and then the quantity thus obtained would be called the *sample* standard deviation. Dividing by $n-1$ gives a better estimate of the population standard deviation than dividing by n .

As a slightly more complicated real-life example, the average height for adult men in the United States is about 70 inches, with a standard deviation of around 3 inches. This means that most men (about 68 percent, assuming a normal distribution) have a height within 3 inches of the mean (67–73 inches) – one standard deviation – and almost all men (about 95%) have a height within 6 inches of the mean (64–76 inches) – two standard deviations. If the standard deviation were zero, then all men would be exactly 70 inches tall. If the standard deviation were 20 inches, then men would have much more variable heights, with a typical range of about 50–90 inches. Three standard deviations account for 99.7 percent of the sample population being studied, assuming the distribution is normal (bell-shaped).

Definition of population values

Let X be a random variable with mean value μ :

$$E[X] = \mu.$$

Here the operator E denotes the average or expected value of X . Then the **standard deviation** of X is the quantity

$$\begin{aligned} \sigma &= \sqrt{E[(X - \mu)^2]} \\ &= \sqrt{E[X^2] + E[(-2\mu X)] + E[\mu^2]} = \sqrt{E[X^2] - 2\mu E[X] + \mu^2} \\ &= \sqrt{E[X^2] - 2\mu^2 + \mu^2} = \sqrt{E[X^2] - \mu^2} \\ &= \sqrt{E[X^2] - (E[X])^2} \end{aligned}$$

(derived using the properties of expected value).

In other words the standard deviation σ (sigma) is the square root of the variance of X ; i.e., it is the square root of the average value of $(X - \mu)^2$.

The standard deviation of a (univariate) probability distribution is the same as that of a random variable having that distribution. Not all random variables have a standard deviation, since these expected values need not exist. For example, the standard deviation of a random variable that follows a Cauchy distribution is undefined because its expected value μ is undefined.

Discrete random variable

In the case where X takes random values from a finite data set x_1, x_2, \dots, x_N , with each value having the same probability, the standard deviation is

$$\sigma = \sqrt{\frac{1}{N} [(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2]}, \quad \text{where } \mu = \frac{1}{N}(x_1 + \dots + x_N),$$

or, using summation notation,

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}, \quad \text{where } \mu = \frac{1}{N} \sum_{i=1}^N x_i.$$

If, instead of having equal probabilities, the values have different probabilities, let x_1 have probability p_1 , x_2 have probability p_2 , ..., x_N have probability p_N . In this case, the standard deviation will be

$$\sigma = \sqrt{\sum_{i=1}^N p_i (x_i - \mu)^2}, \quad \text{where } \mu = \sum_{i=1}^N p_i x_i.$$

Continuous random variable

The standard deviation of a continuous real-valued random variable X with probability density function $p(x)$ is

$$\sigma = \sqrt{\int_{\mathbf{X}} (x - \mu)^2 p(x) dx}, \quad \text{where } \mu = \int_{\mathbf{X}} x p(x) dx,$$

and where the integrals are definite integrals taken for x ranging over the set of possible values of the random variable X .

In the case of a parametric family of distributions, the standard deviation can be expressed in terms of the parameters. For example, in the case of the log-normal distribution with parameters μ and σ^2 , the standard deviation is $[(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)]^{1/2}$.

Estimation

One can find the standard deviation of an entire population in cases (such as standardized testing) where every member of a population is sampled. In cases where that cannot be done, the standard deviation σ is estimated by examining a random sample taken from the population and computing a statistic of the sample, which is used as an estimate of the population standard deviation. Such a statistic is called an estimator, and the estimator (or the value of the estimator, namely the estimate) is called a **sample standard deviation**, and is denoted by s (possibly with modifiers). However, unlike in the case of estimating the population mean, for which the sample mean is a simple estimator with many desirable properties (unbiased, efficient, maximum likelihood), there is no single estimator for the standard deviation with all these properties, and unbiased estimation of standard deviation is a very technical involved problem. Most often, the standard deviation is estimated using the *corrected sample standard deviation* (using $N - 1$), defined below, and this is often referred to as the "sample standard deviation", without qualifiers. However, other estimators are better in other respects: the uncorrected estimator (using N) yields lower mean squared error, while using $N - 1.5$ (for the normal distribution) almost completely eliminates bias.

Uncorrected sample standard deviation

Firstly, the formula for the *population* standard deviation (of a finite population) can be applied to the sample, using the size of the sample as the size of the population (though the actual population size from which the sample is drawn may be much larger). This estimator, denoted by s_N , is known as the **uncorrected sample standard deviation**, or sometimes the **standard deviation of the sample** (considered as the entire population), and is defined as follows:^[citation needed]

$$s_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2},$$

where $\{x_1, x_2, \dots, x_N\}$ are the observed values of the sample items and \bar{x} is the mean value of these observations, while the denominator N stands for the size of the sample.

This is a consistent estimator (it converges in probability to the population value as the number of samples goes to infinity), and is the maximum-likelihood estimate when the population is normally distributed.^[citation needed] However, this is a biased estimator, as the estimates are generally too low. The bias decreases as sample size grows, dropping off as $1/n$, and thus is most significant for small or moderate sample sizes; for $n > 75$ the bias is below 1%. Thus for very large sample sizes, the uncorrected sample standard deviation is generally acceptable. This estimator also has a uniformly smaller mean squared error than the corrected sample standard deviation.

Corrected sample standard deviation

When discussing the bias, to be more precise, the corresponding estimator for the variance, the *biased sample variance*:

$$s_N^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2,$$

equivalently the second central moment of the sample (as the mean is the first moment), is a biased estimator of the variance (it underestimates the population variance). Taking the square root to pass to the standard deviation introduces further downward bias, by Jensen's inequality, due to the square root being a concave function. The bias in the variance is easily corrected, but the bias from the square root is more difficult to correct, and depends on the distribution in question.

An unbiased estimator for the *variance* is given by applying Bessel's correction, using $N - 1$ instead of N to yield the *unbiased sample variance*, denoted s^2 :

$$s^2 = \frac{1}{N - 1} \sum_{i=1}^N (x_i - \bar{x})^2.$$

This estimator is unbiased if the variance exists and the sample values are drawn independently with replacement. $N - 1$ corresponds to the number of degrees of freedom in the vector of residuals, $(x_1 - \bar{x}, \dots, x_n - \bar{x})$.

Taking square roots reintroduces bias, and yields the **corrected sample standard deviation**, denoted by s :

$$s = \sqrt{\frac{1}{N - 1} \sum_{i=1}^N (x_i - \bar{x})^2}.$$

While s^2 is an unbiased estimator for the population variance, s is a biased estimator for the population standard deviation, though markedly less biased than the uncorrected sample standard deviation. The bias is still significant for small samples (n less than 10), and also drops off as $1/n$ as sample size increases. This estimator is commonly used, and generally known simply as the "sample standard deviation".

Unbiased sample standard deviation

For unbiased estimation of standard deviation, there is no formula that works across all distributions, unlike for mean and variance. Instead, s is used as a basis, and is scaled by a correction factor to produce an unbiased estimate. For the normal distribution, an unbiased estimator is given by s/c_4 , where the correction factor (which depends on N) is given in terms of the Gamma function, and equals:

$$c_4(N) = \sqrt{\frac{2}{N-1}} \frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}.$$

This arises because the sampling distribution of the sample standard deviation follows a (scaled) chi distribution, and the correction factor is the mean of the chi distribution.

An approximation is given by replacing $N-1$ with $N-1.5$, yielding:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1.5} \sum_{i=1}^n (x_i - \bar{x})^2},$$

The error in this approximation decays quadratically (as $1/N^2$), and it is suited for all but the smallest samples or highest precision: for $n=3$ the bias is equal to 1.3%, and for $n=9$ the bias is already less than 0.1%.

For other distributions, the correct formula depends on the distribution, but a rule of thumb is to use the further refinement of the approximation:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1.5-\frac{1}{4}\gamma_2} \sum_{i=1}^n (x_i - \bar{x})^2},$$

where γ_2 denotes the population excess kurtosis. The excess kurtosis may be either known beforehand for certain distributions, or estimated from the data.

Confidence interval of a sampled standard deviation

The standard deviation we obtain by sampling a distribution is itself not absolutely accurate, both for mathematical reasons (explained here by the confidence interval) and for practical reasons of measurement (measurement error). The mathematical effect can be described by the confidence interval or CI. To show how a larger sample will increase the confidence interval, consider the following examples: For a small population of $N=2$, the 95% CI of the SD is from $0.45*SD$ to $31.9*SD$. In other words, the standard deviation of the distribution in 95% of the cases can be larger by a factor of 31 or smaller by a factor of 2. For a larger population of $N=10$, the CI is $0.69*SD$ to $1.83*SD$. So even with a sample population of 10, the actual SD can still be almost a factor 2 higher than the sampled SD. For a sample population $N=100$, this is down to $0.88*SD$ to $1.16*SD$. To be more certain that the sampled SD is close to the actual SD we need to sample a large number of points.

Identities and mathematical properties

The standard deviation is invariant under changes in location, and scales directly with the scale of the random variable. Thus, for a constant c and random variables X and Y :

$$\begin{aligned}\sigma(c) &= 0 \\ \sigma(X+c) &= \sigma(X), \\ \sigma(cX) &= |c|\sigma(X).\end{aligned}$$

The standard deviation of the sum of two random variables can be related to their individual standard deviations and the covariance between them:

$$\sigma(X+Y) = \sqrt{\text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)}.$$

where $\text{var} = \sigma^2$ and cov stand for variance and covariance, respectively.

The calculation of the sum of squared deviations can be related to moments calculated directly from the data. The standard deviation of the sample can be computed as:

$$\sigma(X) = \sqrt{E[(X - E(X))^2]} = \sqrt{E[X^2] - (E[X])^2}.$$

The sample standard deviation can be computed as:

$$\sigma(X) = \sqrt{\frac{N}{N-1}} \sqrt{E[(X - E(X))^2]}.$$

For a finite population with equal probabilities at all points, we have

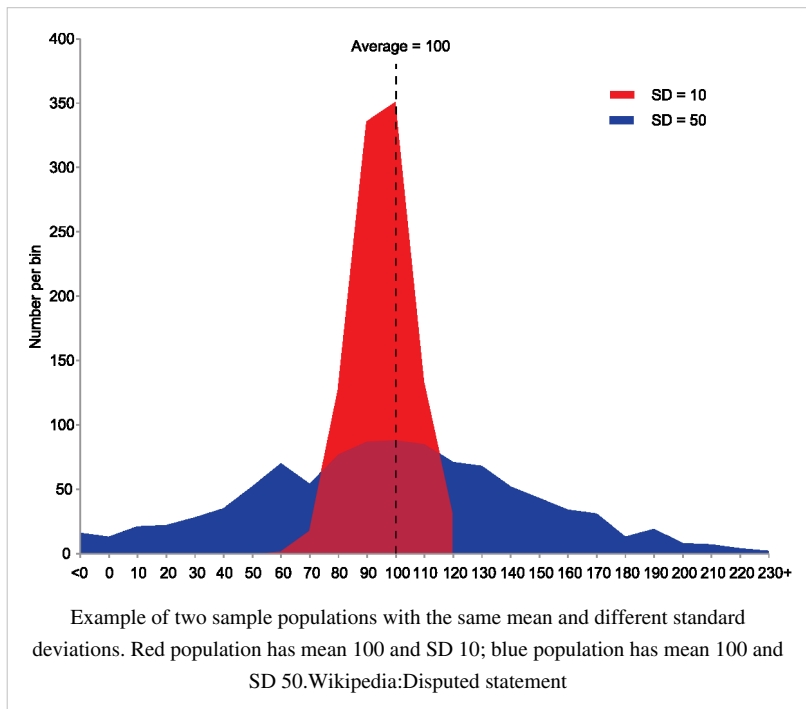
$$\sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} = \sqrt{\frac{1}{N} \left(\sum_{i=1}^N x_i^2 \right) - \bar{x}^2} = \sqrt{\left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \left(\frac{1}{N} \sum_{i=1}^N x_i \right)^2}.$$

This means that the standard deviation is equal to the square root of (the average of the squares less the square of the average). See computational formula for the variance for proof, and for an analogous result for the sample standard deviation.

Interpretation and application

A large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

For example, each of the three populations $\{0, 0, 14, 14\}$, $\{0, 6, 8, 14\}$ and $\{6, 6, 8, 8\}$ has a mean of 7. Their standard deviations are 7, 5, and 1, respectively. The third population has a much smaller standard deviation than the other two because its values are all close to 7. It will have the same units as the data points themselves. If, for instance, the data set $\{0, 6, 8, 14\}$ represents the ages of a population of four siblings in years, the standard deviation is 5 years. As another



example, the population $\{1000, 1006, 1008, 1014\}$ may represent the distances traveled by four athletes, measured in meters. It has a mean of 1007 meters, and a standard deviation of 5 meters.

Standard deviation may serve as a measure of uncertainty. In physical science, for example, the reported standard deviation of a group of repeated measurements gives the precision of those measurements. When deciding whether measurements agree with a theoretical prediction, the standard deviation of those measurements is of crucial importance: if the mean of the measurements is too far away from the prediction (with the distance measured in standard deviations), then the theory being tested probably needs to be revised. This makes sense since they fall outside the range of values that could reasonably be expected to occur, if the prediction were correct and the standard deviation appropriately quantified. See prediction interval.

While the standard deviation does measure how far typical values tend to be from the mean, other measures are available. An example is the mean absolute deviation, which might be considered a more direct measure of average distance, compared to the root mean square distance inherent in the standard deviation.

Application examples

The practical value of understanding the standard deviation of a set of values is in appreciating how much variation there is from the average (mean).

Climate

As a simple example, consider the average daily maximum temperatures for two cities, one inland and one on the coast. It is helpful to understand that the range of daily maximum temperatures for cities near the coast is smaller than for cities inland. Thus, while these two cities may each have the same average maximum temperature, the standard deviation of the daily maximum temperature for the coastal city will be less than that of the inland city as, on any particular day, the actual maximum temperature is more likely to be farther from the average maximum temperature for the inland city than for the coastal one.

Particle physics

Particle physics uses a standard of "5 sigma" for the declaration of a discovery. At five-sigma there is only one chance in nearly two million that a random fluctuation would yield the result. This level of certainty prompted the announcement that a particle consistent with the Higgs boson has been discovered in two independent experiments at CERN.

Finance

In finance, standard deviation is often used as a measure of the risk associated with price-fluctuations of a given asset (stocks, bonds, property, etc.), or the risk of a portfolio of assets (actively managed mutual funds, index mutual funds, or ETFs). Risk is an important factor in determining how to efficiently manage a portfolio of investments because it determines the variation in returns on the asset and/or portfolio and gives investors a mathematical basis for investment decisions (known as mean-variance optimization). The fundamental concept of risk is that as it increases, the expected return on an investment should increase as well, an increase known as the risk premium. In other words, investors should expect a higher return on an investment when that investment carries a higher level of risk or uncertainty. When evaluating investments, investors should estimate both the expected return and the uncertainty of future returns. Standard deviation provides a quantified estimate of the uncertainty of future returns.

For example, let's assume an investor had to choose between two stocks. Stock A over the past 20 years had an average return of 10 percent, with a standard deviation of 20 percentage points (pp) and Stock B, over the same period, had average returns of 12 percent but a higher standard deviation of 30 pp. On the basis of risk and return, an investor may decide that Stock A is the safer choice, because Stock B's additional two percentage points of return is not worth the additional 10 pp standard deviation (greater risk or uncertainty of the expected return). Stock B is likely to fall short of the initial investment (but also to exceed the initial investment) more often than Stock A under the same circumstances, and is estimated to return only two percent more on average. In this example, Stock A is expected to earn about 10 percent, plus or minus 20 pp (a range of 30 percent to -10 percent), about two-thirds of the future year returns. When considering more extreme possible returns or outcomes in future, an investor should expect results of as much as 10 percent plus or minus 60 pp, or a range from 70 percent to -50 percent, which includes outcomes for three standard deviations from the average return (about 99.7 percent of probable returns).

Calculating the average (or arithmetic mean) of the return of a security over a given period will generate the expected return of the asset. For each period, subtracting the expected return from the actual return results in the difference from the mean. Squaring the difference in each period and taking the average gives the overall variance of the return of the asset. The larger the variance, the greater risk the security carries. Finding the square root of this

variance will give the standard deviation of the investment tool in question.

Population standard deviation is used to set the width of Bollinger Bands, a widely adopted technical analysis tool. For example, the upper Bollinger Band is given as $x + n\sigma_x$. The most commonly used value for n is 2; there is about a five percent chance of going outside, assuming a normal distribution of returns.

Unfortunately, financial time series are known to be non-stationary series, whereas the statistical calculations above, such as standard deviation, apply only to stationary series. Whatever apparent "predictive powers" or "forecasting ability" that may appear when applied as above is illusory. To apply the above statistical tools to non-stationary series, the series first must be transformed to a stationary series, enabling use of statistical tools that now have a valid basis from which to work.

Geometric interpretation



It is requested that a **diagram** or **diagrams** be included in this article to improve its quality. Specific illustrations, plots or diagrams can be requested at the Graphic Lab.

For more information, refer to discussion on this page and/or the listing at Wikipedia:Requested images.

To gain some geometric insights and clarification, we will start with a population of three values, x_1, x_2, x_3 . This defines a point $P = (x_1, x_2, x_3)$ in \mathbf{R}^3 . Consider the line $L = \{(r, r, r) : r \in \mathbf{R}\}$. This is the "main diagonal" going through the origin. If our three given values were all equal, then the standard deviation would be zero and P would lie on L . So it is not unreasonable to assume that the standard deviation is related to the *distance* of P to L . And that is indeed the case. To move orthogonally from L to the point P , one begins at the point:

$$M = (\bar{x}, \bar{x}, \bar{x})$$

whose coordinates are the mean of the values we started out with. A little algebra shows that the distance between P and M (which is the same as the orthogonal distance between P and the line L) is equal to the standard deviation of the vector x_1, x_2, x_3 , multiplied by the square root of the number of dimensions of the vector (3 in this case.)

Chebyshev's inequality

An observation is rarely more than a few standard deviations away from the mean. Chebyshev's inequality ensures that, for all distributions for which the standard deviation is defined, the amount of data within a number of standard deviations of the mean is at least as much as given in the following table.

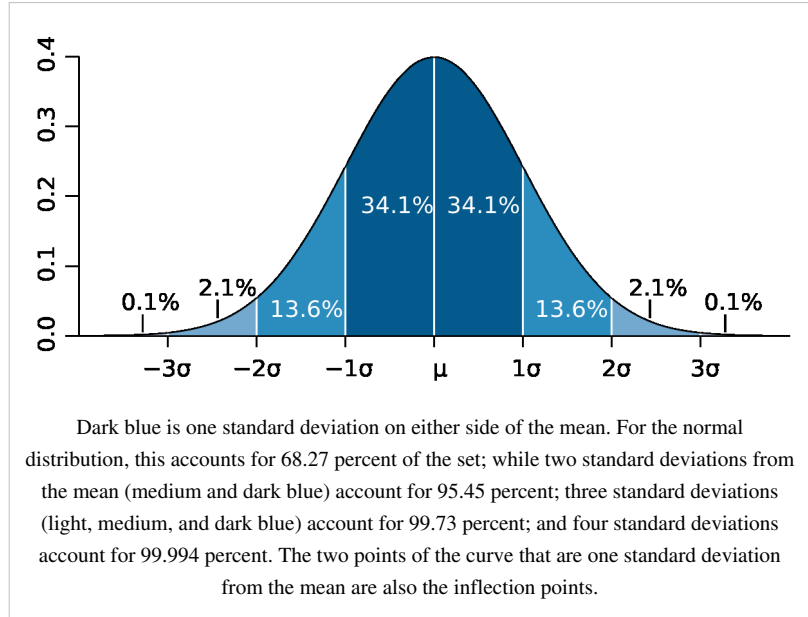
Minimum population	Distance from mean
50%	$\sqrt{2}$
75%	2
89%	3
94%	4
96%	5
97%	6
$1 - \frac{1}{k^2}$ [1]	k
i	$\frac{1}{\sqrt{i-1}}$

Rules for normally distributed data

The central limit theorem says that the distribution of an average of many independent, identically distributed random variables tends toward the famous bell-shaped normal distribution with a probability density function of:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where μ is the expected value of the random variables, σ equals their distribution's standard deviation divided by $n^{1/2}$, and n is the number of random variables. The standard deviation therefore is simply a scaling variable that adjusts how broad the curve will be, though it also appears in the normalizing constant.



If a data distribution is approximately normal, then the proportion of data values within z standard deviations of the mean is defined by:

$$\text{Proportion} = \text{erf}\left(\frac{z}{\sqrt{2}}\right)$$

where erf is the error function. If a data distribution is approximately normal then about 68 percent of the data values are within one standard deviation of the mean (mathematically, $\mu \pm \sigma$, where μ is the arithmetic mean), about 95 percent are within two standard deviations ($\mu \pm 2\sigma$), and about 99.7 percent lie within three standard deviations ($\mu \pm 3\sigma$). This is known as the *68-95-99.7 rule*, or *the empirical rule*.

For various values of z , the percentage of values expected to lie in and outside the symmetric interval, $\text{CI} = (-z\sigma, z\sigma)$, are as follows:

$z\sigma$	Percentage within CI	Percentage outside CI	Fraction outside CI
0.674490σ	50%	50%	1 / 2
0.994458σ	68%	32%	1 / 3.125
1σ	68.2689492%	31.7310508%	1 / 3.1514872
1.281552σ	80%	20%	1 / 5
1.644854σ	90%	10%	1 / 10
1.959964σ	95%	5%	1 / 20
2σ	95.4499736%	4.5500264%	1 / 21.977895
2.575829σ	99%	1%	1 / 100
3σ	99.7300204%	0.2699796%	1 / 370.398
3.290527σ	99.9%	0.1%	1 / 1000
3.890592σ	99.99%	0.01%	1 / 10000
4σ	99.993666%	0.006334%	1 / 15787
4.417173σ	99.999%	0.001%	1 / 100000

4.5σ	99.9993204653751%	0.0006795346249%	3.4 / 1000000 (on each side of mean)
4.891638σ	99.9999%	0.0001%	1 / 1000000
5σ	99.9999426697%	0.0000573303%	1 / 1744278
5.326724σ	99.99999%	0.00001%	1 / 10000000
5.730729σ	99.999999%	0.000001%	1 / 100000000
6σ	99.999998027%	0.000001973%	1 / 506797346
6.109410σ	99.9999999%	0.0000001%	1 / 1000000000
6.466951σ	99.99999999%	0.00000001%	1 / 10000000000
6.806502σ	99.999999999%	0.000000001%	1 / 100000000000
7σ	99.999999997440%	0.00000000256%	1 / 390682215445

Relationship between standard deviation and mean

The mean and the standard deviation of a set of data are descriptive statistics usually reported together. In a certain sense, the standard deviation is a "natural" measure of statistical dispersion if the center of the data is measured about the mean. This is because the standard deviation from the mean is smaller than from any other point. The precise statement is the following: suppose x_1, \dots, x_n are real numbers and define the function:

$$\sigma(r) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - r)^2}.$$

Using calculus or by completing the square, it is possible to show that $\sigma(r)$ has a unique minimum at the mean:

$$r = \bar{x}.$$

Variability can also be measured by the coefficient of variation, which is the ratio of the standard deviation to the mean. It is a dimensionless number.

Standard deviation of the mean

Often, we want some information about the precision of the mean we obtained. We can obtain this by determining the standard deviation of the sampled mean. Assuming statistical independence of the values in the sample, the standard deviation of the mean is related to the standard deviation of the distribution by:

$$\sigma_{\text{mean}} = \frac{1}{\sqrt{N}} \sigma$$

where N is the number of observations in the sample used to estimate the mean. This can easily be proven with (see basic properties of the variance):

$$\begin{aligned} \text{var}(X) &\equiv \sigma_X^2 \\ \text{var}(X_1 + X_2) &\equiv \text{var}(X_1) + \text{var}(X_2) \\ \text{var}(cX_1) &\equiv c^2 \text{var}(X_1) \end{aligned}$$

hence

$$\begin{aligned} \text{var}(\text{mean}) &= \text{var}\left(\frac{1}{N} \sum_{i=1}^N X_i\right) = \frac{1}{N^2} \text{var}\left(\sum_{i=1}^N X_i\right) \\ &= \frac{1}{N^2} \sum_{i=1}^N \text{var}(X_i) = \frac{N}{N^2} \text{var}(X) = \frac{1}{N} \text{var}(X). \end{aligned}$$

Resulting in:

$$\sigma_{\text{mean}} = \frac{\sigma}{\sqrt{N}}.$$

It should be emphasized that in order to estimate standard deviation of the mean σ_{mean} it is necessary to know standard deviation of the entire population σ beforehand. However, in most applications this parameter is unknown. For example, if series of 10 measurements of previously unknown quantity is performed in laboratory, it is possible to calculate resulting sample mean and sample standard deviation, but it is impossible to calculate standard deviation of the mean.

Rapid calculation methods

The following two formulas can represent a running (repeatedly updated) standard deviation. A set of two power sums s_1 and s_2 are computed over a set of N values of x , denoted as x_1, \dots, x_N :

$$s_j = \sum_{k=1}^N x_k^j.$$

Given the results of these running summations, the values N, s_1, s_2 can be used at any time to compute the *current* value of the running standard deviation:

$$\sigma = \frac{\sqrt{N s_2 - s_1^2}}{N}$$

Where : $N = s_0 = \sum_{k=1}^N x_k^0$.

Similarly for sample standard deviation,

$$s = \sqrt{\frac{N s_2 - s_1^2}{N(N-1)}}.$$

In a computer implementation, as the three s_j sums become large, we need to consider round-off error, arithmetic overflow, and arithmetic underflow. The method below calculates the running sums method with reduced rounding errors. This is a "one pass" algorithm for calculating variance of n samples without the need to store prior data during the calculation. Applying this method to a time series will result in successive values of standard deviation corresponding to n data points as n grows larger with each new sample, rather than a constant-width sliding window calculation.

For $k = 1, \dots, n$:

$$A_0 = 0$$

$$A_k = A_{k-1} + \frac{x_k - A_{k-1}}{k}$$

where A is the mean value.

$$Q_0 = 0$$

$$Q_k = Q_{k-1} + \frac{k-1}{k} (x_k - A_{k-1})^2 = Q_{k-1} + (x_k - A_{k-1})(x_k - A_k)$$

Note: $Q_1 = 0$ since $k-1 = 0$ or $x_1 = A_1$

Sample variance:

$$s_n^2 = \frac{Q_n}{n-1}$$

Population variance:

$$\sigma_n^2 = \frac{Q_n}{n}$$

Weighted calculation

When the values x_i are weighted with unequal weights w_i , the power sums s_0, s_1, s_2 are each computed as:

$$s_j = \sum_{k=1}^N w_k x_k^j.$$

And the standard deviation equations remain unchanged. Note that s_0 is now the sum of the weights and not the number of samples N .

The incremental method with reduced rounding errors can also be applied, with some additional complexity.

A running sum of weights must be computed for each k from 1 to n :

$$W_0 = 0$$

$$W_k = W_{k-1} + w_k$$

and places where $1/n$ is used above must be replaced by w_i/W_n :

$$A_0 = 0$$

$$A_k = A_{k-1} + \frac{w_k}{W_k}(x_k - A_{k-1})$$

$$Q_0 = 0$$

$$Q_k = Q_{k-1} + \frac{w_k W_{k-1}}{W_k}(x_k - A_{k-1})^2 = Q_{k-1} + w_k(x_k - A_{k-1})(x_k - A_k)$$

In the final division,

$$\sigma_n^2 = \frac{Q_n}{W_n}$$

and

$$s_n^2 = \frac{n'}{n' - 1} \sigma_n^2$$

where n is the total number of elements, and n' is the number of elements with non-zero weights. The above formulas become equal to the simpler formulas given above if weights are taken as equal to one.

Combining standard deviations

Population-based statistics

The populations of sets, which may overlap, can be calculated simply as follows:

$$N_{X \cup Y} = N_X + N_Y - N_{X \cap Y}$$

$$X \cap Y = \emptyset \Rightarrow N_{X \cap Y} = 0$$

$$\Rightarrow N_{X \cup Y} = N_X + N_Y$$

Standard deviations of non-overlapping ($X \cap Y = \emptyset$) sub-populations can be aggregated as follows if the size (actual or relative to one another) and means of each are known:

$$\mu_{X \cup Y} = \frac{N_X \mu_X + N_Y \mu_Y}{N_X + N_Y}$$

$$\sigma_{X \cup Y} = \sqrt{\frac{N_X \sigma_X^2 + N_Y \sigma_Y^2}{N_X + N_Y} + \frac{N_X N_Y}{(N_X + N_Y)^2} (\mu_X - \mu_Y)^2}$$

For example, suppose it is known that the average American man has a mean height of 70 inches with a standard deviation of three inches and that the average American woman has a mean height of 65 inches with a standard deviation of two inches. Also assume that the number of men, N , is equal to the number of women. Then the mean and standard deviation of heights of American adults could be calculated as:

$$\mu = \frac{N \cdot 70 + N \cdot 65}{N + N} = \frac{70 + 65}{2} = 67.5$$

$$\sigma = \sqrt{\frac{3^2 + 2^2}{2} + \frac{(70 - 65)^2}{2^2}} = \sqrt{12.75} \approx 3.57$$

For the more general case of M non-overlapping populations, X_1 through X_M , and the aggregate population $X = \bigcup_i X_i$:

$$\mu_X = \frac{\sum_i N_{X_i} \mu_{X_i}}{\sum_i N_{X_i}}$$

$$\sigma_X = \sqrt{\frac{\sum_i N_{X_i} (\sigma_{X_i}^2 + \mu_{X_i}^2)}{\sum_i N_{X_i}} - \mu_X^2} = \sqrt{\frac{\sum_i N_{X_i} \sigma_{X_i}^2}{\sum_i N_{X_i}} + \frac{\sum_{i < j} N_{X_i} N_{X_j} (\mu_{X_i} - \mu_{X_j})^2}{(\sum_i N_{X_i})^2}}$$

where

$$X_i \cap X_j = \emptyset, \quad \forall i < j.$$

If the size (actual or relative to one another), mean, and standard deviation of two overlapping populations are known for the populations as well as their intersection, then the standard deviation of the overall population can still be calculated as follows:

$$\mu_{X \cup Y} = \frac{1}{N_{X \cup Y}} (N_X \mu_X + N_Y \mu_Y - N_{X \cap Y} \mu_{X \cap Y})$$

$$\sigma_{X \cup Y} = \sqrt{\frac{1}{N_{X \cup Y}} (N_X [\sigma_X^2 + \mu_X^2] + N_Y [\sigma_Y^2 + \mu_Y^2] - N_{X \cap Y} [\sigma_{X \cap Y}^2 + \mu_{X \cap Y}^2]) - \mu_{X \cup Y}^2}$$

If two or more sets of data are being added together datapoint by datapoint, the standard deviation of the result can be calculated if the standard deviation of each data set and the covariance between each pair of data sets is known:

$$\sigma_X = \sqrt{\sum_i \sigma_{X_i}^2 + \sum_{i,j} \text{cov}(X_i, X_j)}$$

For the special case where no correlation exists between any pair of data sets, then the relation reduces to the root-mean-square:

$$\text{cov}(X_i, X_j) = 0, \quad \forall i < j$$

$$\Rightarrow \sigma_X = \sqrt{\sum_i \sigma_{X_i}^2}.$$

Sample-based statistics

Standard deviations of non-overlapping ($X \cap Y = \emptyset$) sub-samples can be aggregated as follows if the actual size and means of each are known:

$$\mu_{X \cup Y} = \frac{1}{N_{X \cup Y}} (N_X \mu_X + N_Y \mu_Y)$$

$$\sigma_{X \cup Y} = \sqrt{\frac{1}{N_{X \cup Y} - 1} ([N_X - 1] \sigma_X^2 + N_X \mu_X^2 + [N_Y - 1] \sigma_Y^2 + N_Y \mu_Y^2 - [N_X + N_Y] \mu_{X \cup Y}^2)}$$

For the more general case of M non-overlapping data sets, X_1 through X_M , and the aggregate data set $X = \bigcup_i X_i$:

$$\mu_X = \frac{1}{\sum_i N_{X_i}} \left(\sum_i N_{X_i} \mu_{X_i} \right)$$

$$\sigma_X = \sqrt{\frac{1}{\sum_i N_{X_i} - 1} \left(\sum_i [(N_{X_i} - 1)\sigma_{X_i}^2 + N_{X_i} \mu_{X_i}^2] - \left[\sum_i N_{X_i} \right] \mu_X^2 \right)}$$

where:

$$X_i \cap X_j = \emptyset, \quad \forall i < j.$$

If the size, mean, and standard deviation of two overlapping samples are known for the samples as well as their intersection, then the standard deviation of the aggregated sample can still be calculated. In general:

$$\mu_{X \cup Y} = \frac{1}{N_{X \cup Y}} (N_X \mu_X + N_Y \mu_Y - N_{X \cap Y} \mu_{X \cap Y})$$

$$\sigma_{X \cup Y} = \sqrt{\frac{[N_X - 1]\sigma_X^2 + N_X \mu_X^2 + [N_Y - 1]\sigma_Y^2 + N_Y \mu_Y^2 - [N_{X \cap Y} - 1]\sigma_{X \cap Y}^2 - N_{X \cap Y} \mu_{X \cap Y}^2 - [N_X + N_Y - N_{X \cap Y}]\mu_{X \cup Y}^2}{N_{X \cup Y} - 1}}$$

History

The term *standard deviation* was first used in writing by Karl Pearson in 1894, following his use of it in lectures. This was as a replacement for earlier alternative names for the same idea: for example, Gauss used *mean error*. It may be worth noting in passing that the mean error is mathematically distinct from the standard deviation.

References

[1] Ghahramani, Saeed (2000). *Fundamentals of Probability* (2nd Edition). Prentice Hall: New Jersey. p. 438.

External links

- Hazewinkel, Michiel, ed. (2001), "Quadratic deviation" (<http://www.encyclopediaofmath.org/index.php?title=p/q076030>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- A simple way to understand Standard Deviation (<http://standard-deviation.appspot.com/>)
- Standard Deviation – an explanation without maths (<http://www.techbookreport.com/tutorials/stddev-30-secs.html>)
- Standard Deviation, an elementary introduction (<http://davidmlane.com/hyperstat/A16252.html>)
- Standard Deviation while Financial Modeling in Excel (<http://www.edupristine.com/blog/what-is-standard-deviation>)
- Standard Deviation, a simpler explanation for writers and journalists (<http://www.robertniles.com/stats/stdev.shtml>)
- The concept of Standard Deviation is shown in this 8-foot-tall (2.4 m) Probability Machine (named Sir Francis) comparing stock market returns to the randomness of the beans dropping through the quincunx pattern. (<http://www.youtube.com/watch?v=AUSKtk9ENzg>) from Index Funds Advisors IFA.com (<http://www.ifa.com>)

Coefficient of variation

In probability theory and statistics, the **coefficient of variation (CV)** is a normalized measure of dispersion of a probability distribution or frequency distribution. It is also known as **unitized risk** or the **variation coefficient**. The absolute value of the CV is sometimes known as relative standard deviation (RSD), which is expressed as a percentage.

Definition

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ :

$$c_v = \frac{\sigma}{\mu}$$

which is the inverse of one definition of the signal-to-noise ratio. It shows the extent of variability in relation to mean of the population.

The coefficient of variation should be computed only for data measured on a ratio scale, as these are measurements that can only take non-negative values. The coefficient of variation may not have any meaning for data on an interval scale. For example, most temperature scales are interval scales (e.g., Celsius, Fahrenheit etc.) that can take both positive and negative values, whereas the Kelvin scale has an absolute null value (i.e., 0K is the absence of heat), and negative values are nonsensical. Hence, the Kelvin scale is a ratio scale. While the standard deviation (SD) can be derived on both the Kelvin and the Celsius scale (with both leading to the same SDs), the CV is only relevant as a measure of relative variability for the Kelvin scale.

Often, laboratory values that are measured based on chromatographic methods are log-normally distributed. In this case, the CV would be constant over a large range of measurements, while SDs would vary depending on typical values that are being measured.

A nonparametric possibility is the quartile coefficient of dispersion, i.e. interquartile range $Q_3 - Q_1$ divided by the median Q_2 .

Estimation

When only a sample of data from a population is available, the population CV can be estimated using the ratio of the sample standard deviation s to the sample mean \bar{x} :

$$\hat{c}_v = \frac{s}{\bar{x}}$$

But this estimator, when applied to a small or moderately sized sample, tends to be too low: it is a biased estimator. For normally distributed data, an unbiased estimator^[1] for a sample of size n is:

$$\hat{c}_v^* = \left(1 + \frac{1}{4n}\right) \hat{c}_v$$

In many applications, it can be assumed that data are log-normally distributed (evidenced by the presence of skewness in the sampled data). In such cases, a more accurate estimate, derived from the properties of the log-normal distribution, is defined as:

$$\hat{c}_{vln} = \sqrt{e^{s_{ln}^2} - 1}$$

where s_{ln} is the sample standard deviation of the data after a natural log transformation. (In the event that measurements are recorded using any other logarithmic base, b , their standard deviation s_b is converted to base e using $s_{ln} = s_b \ln(b)$, and the formula for \hat{c}_{vln} remains the same.^[2]) This estimate is sometimes referred to as the "geometric coefficient of variation"^[3] in order to distinguish it from the simple estimate above. However, "geometric coefficient of variation" has also been defined as:

$$GCV = e^{s_{ln}} - 1$$

This term was intended to be *analogous* to the coefficient of variation, for describing multiplicative variation in log-normal data, but this definition of GCV has no theoretical basis as an estimate of c_v itself.

For many practical purposes (such as sample size determination and calculation of confidence intervals) it is s_{ln} which is of most use in the context of log-normally distributed data. If necessary, this can be derived from an estimate of c_v or GCV by inverting the corresponding formula.

Laboratory measures of intra and inter-assay CVs

CV measures are often used as quality controls for quantitative laboratory assays. While intra-assay and inter-assay CVs might be assumed to be calculated by simply averaging CV values across CV values for multiple samples within one assay or by averaging multiple inter-assay CV estimates, it has been suggested that these practices are incorrect and that a more complex computational process is required.

Comparison to standard deviation

Advantages

The coefficient of variation is useful because the standard deviation of data must always be understood in the context of the mean of the data. In contrast, the actual value of the CV is independent of the unit in which the measurement has been taken, so it is a dimensionless number. For comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation.

Disadvantages

- When the mean value is close to zero, the coefficient of variation will approach infinity and is therefore sensitive to small changes in the mean. This is often the case if the values do not originate from a ratio scale.
- Unlike the standard deviation, it cannot be used directly to construct confidence intervals for the mean.

Applications

The coefficient of variation is also common in applied probability fields such as renewal theory, queueing theory, and reliability theory. In these fields, the exponential distribution is often more important than the normal distribution. The standard deviation of an exponential distribution is equal to its mean, so its coefficient of variation is equal to 1. Distributions with $CV < 1$ (such as an Erlang distribution) are considered low-variance, while those with $CV > 1$ (such as a hyper-exponential distribution) are considered high-variance. Some formulas in these fields are expressed using the **squared coefficient of variation**, often abbreviated SCV. In modeling, a variation of the CV is the CV(RMSD). Essentially the CV(RMSD) replaces the standard deviation term with the Root Mean Square Deviation (RMSD). While many natural processes indeed show a correlation between the average value and the amount of variation around it, accurate sensor devices need to be designed in such a way that the coefficient of variation is close to zero, i.e., yielding a constant absolute error over their working range.

Distribution

Provided that negative and small positive values of the sample mean occur with negligible frequency, the probability distribution of the coefficient of variation for a sample of size n has been shown by Hendricks and Robey to be

$$dF_{c_v} = \frac{2}{\pi^{1/2}\Gamma\left(\frac{n-1}{2}\right)} e^{-\frac{n}{2\left(\frac{\sigma}{\mu}\right)^2} \frac{c_v^2}{1+c_v^2}} \frac{c_v^{n-2}}{(1+c_v^2)^{n/2}} \sum_{i=0}^{n-1} \frac{(n-1)!\Gamma\left(\frac{n-i}{2}\right)}{(n-1-i)!i!} \frac{n^{i/2}}{2^{i/2}\left(\frac{\sigma}{\mu}\right)^i} \frac{1}{(1+c_v^2)^{i/2}} dc_v,$$

where the symbol \sum' indicates that the summation is over only even values of $n-1-i$, i.e., if n is odd, sum over even values of i and if n is even, sum only over odd values of i .

This is useful, for instance, in the construction of hypothesis tests or confidence intervals. Statistical inference for the coefficient of variation in normally distributed data is often based on McKay's chi-square approximation for the coefficient of variation ^[4]

Similar ratios

Standardized moments are similar ratios, μ_k/σ^k , which are also dimensionless and scale invariant. The variance-to-mean ratio, σ^2/μ , is another similar ratio, but is not dimensionless, and hence not scale invariant. See Normalization (statistics) for further ratios.

In signal processing, particularly image processing, the reciprocal ratio μ/σ is referred to as the signal to noise ratio.

- Relative standard deviation, $|\sigma/\mu|$
- Standardized moment, μ_k/σ^k
- Variance-to-mean ratio, σ^2/μ
- Fano factor, σ_W^2/μ_W (windowed VMR)
- Signal-to-noise ratio, μ/σ (in signal processing)
 - Signal-to-noise ratio (image processing)

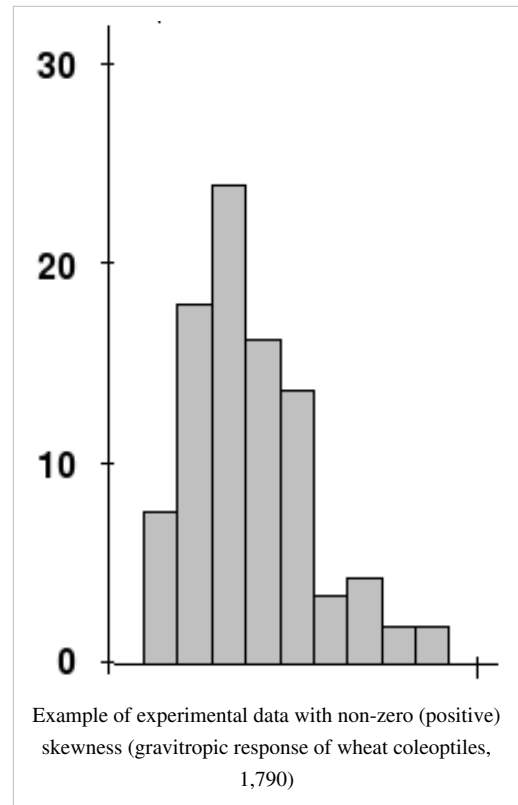
References

- [1] Sokal RR & Rohlf FJ. *Biometry* (3rd Ed). New York: Freeman, 1995. p. 58. ISBN 0-7167-2411-1
- [2] Reed JF, Lynn F, Meade BD. (2002) "Use of Coefficient of Variation in Assessing Variability of Quantitative Assays". *Clin Diagn Lab Immunol.* 9(6): 1235–1239.
- [3] Sawant,S.; Mohan, N. (2011) "FAQ: Issues with Efficacy Analysis of Clinical Trial Data Using SAS" (<http://pharmasug.org/proceedings/2011/PO/PharmaSUG-2011-PO08.pdf>), *PharmaSUG2011*, Paper PO08
- [4] Bennett, B. M. (1976). On an approximate test for homogeneity of coefficients of variation. In: Ziegler, W. J. (ed.) *Contributions to Applied Statistics Dedicated to A. Linder. Experientia Suppl.* 22, 169-171

Skewness

In probability theory and statistics, **skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined.

The qualitative interpretation of the skew is complicated. For a unimodal distribution, negative skew indicates that the *tail* on the left side of the probability density function is longer or fatter than the right side – it does not distinguish these shapes. Conversely, positive skew indicates that the tail on the right side is longer or fatter than the left side. In cases where one tail is long but the other tail is fat, skewness does not obey a simple rule. For example, a zero value indicates that the tails on both sides of the mean balance out, which is the case both for a symmetric distribution, and for asymmetric distributions where the asymmetries even out, such as one tail being long but thin, and the other being short but fat. Further, in multimodal distributions and discrete distributions, skewness is also difficult to interpret. Importantly, the skewness does not determine the relationship of mean and median.



Introduction

Consider the distribution in the figure. The bars on the right side of the distribution taper differently than the bars on the left side. These tapering sides are called *tails*, and they provide a visual means for determining which of the two kinds of skewness a distribution has:

1. *negative skew*: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be *left-skewed*, *left-tailed*, or *skewed to the left*.^[1] Example (observations): 1,1001,1002,1003.
2. *positive skew*: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be *right-skewed*, *right-tailed*, or *skewed to the right*. Example (observations): 1,2,3,1000.

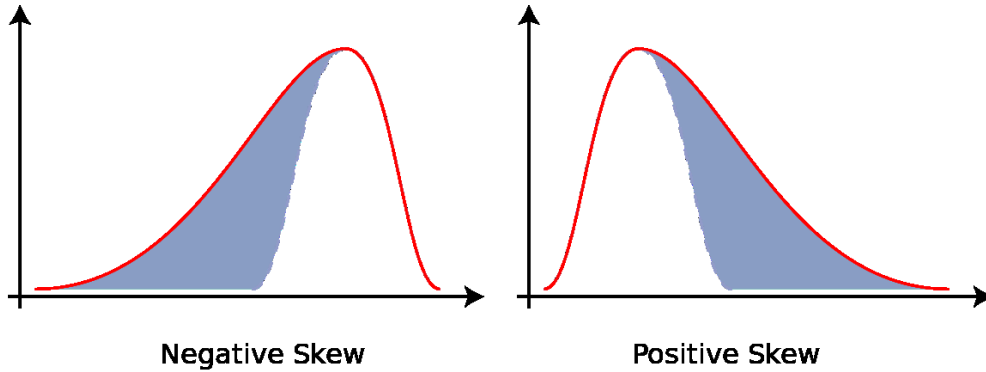
Relationship of mean and median

The skewness is not strictly connected with the relationship between the mean and median: a distribution with negative skew can have the mean greater than or less than the median, and likewise for positive skew.

In the older notion of nonparametric skew, defined as $(\mu - \nu)/\sigma$, where μ is the mean, ν is the median, and σ is the standard deviation, the skewness is defined in terms of this relationship: positive/right nonparametric skew means the mean is greater than (to the right of) the median, while negative/left nonparametric skew means the mean is less than (to the left of) the median. However, the modern definition of skewness and the traditional nonparametric definition do not in general have the same sign: while they agree for some families of distributions, they differ in general, and conflating them is misleading.

If the distribution is symmetric then the mean is equal to the median and the distribution will have zero skewness. If, in addition, the distribution is unimodal, then the mean = median = mode. This is the case of a coin toss or the series 1,2,3,4,... Note, however, that the converse is not true in general, i.e. zero skewness does not imply that the mean is equal to the median.

"Many textbooks," a 2005 article points out, "teach a rule of thumb stating that the mean is right of the median under right skew, and left of the median under left skew. [But] this rule fails with surprising frequency. It can fail in multimodal distributions, or in distributions where one tail is long but the other is fat. Most commonly, though, the rule fails in discrete distributions where the areas to the left and right of the median are not equal. Wikipedia:Please clarify Such distributions not only contradict the textbook relationship between mean, median, and skew, they also contradict the textbook interpretation of the median."



Definition

The skewness of a random variable X is the third standardized moment, denoted γ_1 and defined as

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{E[(X - \mu)^3]}{(E[(X - \mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}},$$

where μ_3 is the third central moment μ , σ is the standard deviation, and E is the expectation operator. The last equality expresses skewness in terms of the ratio of the third cumulant κ_3 and the 1.5th power of the second cumulant κ_2 . This is analogous to the definition of kurtosis as the fourth cumulant normalized by the square of the second cumulant.

The skewness is also sometimes denoted $Skew[X]$.

The formula expressing skewness in terms of the non-central moment $E[X^3]$ can be expressed by expanding the previous formula,

$$\begin{aligned} \gamma_1 &= E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] \\ &= \frac{E[X^3] - 3\mu E[X^2] + 3\mu^2 E[X] - \mu^3}{\sigma^3} \\ &= \frac{E[X^3] - 3\mu(E[X^2] - \mu E[X]) - \mu^3}{\sigma^3} \\ &= \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}. \end{aligned}$$

Sample skewness

For a sample of n values the *sample skewness* is

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}},$$

where \bar{x} is the sample mean, m_3 is the sample third central moment, and m_2 is the sample variance.

Given samples from a population, the equation for the sample skewness g_1 above is a biased estimator of the population skewness. (Note that for a discrete distribution the sample skewness may be undefined (0/0), so its expected value will be undefined.) The usual estimator of population skewness is ^[citation needed]

$$G_1 = \frac{k_3}{k_2^{3/2}} = \frac{\sqrt{n(n-1)}}{n-2} g_1,$$

where k_3 is the unique symmetric unbiased estimator of the third cumulant and k_2 is the symmetric unbiased estimator of the second cumulant. Unfortunately G_1 is, nevertheless, generally biased (although it obviously has the correct expected value of zero for a symmetric distribution). Its expected value can even have the opposite sign from the true skewness. For instance a mixed distribution consisting of very thin Gaussians centred at -99 , 0.5 , and 2 with weights 0.01 , 0.66 , and 0.33 has a skewness of about -9.77 , but in a sample of 3 , G_1 has an expected value of about 0.32 , since usually all three samples are in the positive-valued part of the distribution, which is skewed the other way.

The variance of the skewness of a sample of size n from a normal distribution is ^{[2][3]}

$$\text{var}(G_1) = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}.$$

An approximate alternative is $6/n$ but this is inaccurate for small samples.

Properties

Skewness can be infinite, as when

$$\Pr[X > x] = x^{-3} \text{ for } x > 1, \Pr[X < 1] = 0$$

or undefined, as when

$$\Pr[X < x] = (1-x)^{-3}/2 \text{ for negative } x \text{ and } \Pr[X > x] = (1+x)^{-3}/2 \text{ for positive } x.$$

In this latter example, the third cumulant is undefined. One can also have distributions such as

$$\Pr[X > x] = x^{-2} \text{ for } x > 1, \Pr[X < 1] = 0$$

where both the second and third cumulants are infinite, so the skewness is again undefined.

If Y is the sum of n independent and identically distributed random variables, all with the distribution of X , then the third cumulant of Y is n times that of X and the second cumulant of Y is n times that of X , so $\text{Skew}[Y] = \text{Skew}[X]/\sqrt{n}$. This shows that the skewness of the sum is smaller, as it approaches a Gaussian distribution in accordance with the central limit theorem.

Applications

Skewness has benefits in many areas. Many models assume normal distribution; i.e., data are symmetric about the mean. The normal distribution has a skewness of zero. But in reality, data points may not be perfectly symmetric. So, an understanding of the skewness of the dataset indicates whether deviations from the mean are going to be positive or negative.

D'Agostino's K-squared test is a goodness-of-fit normality test based on sample skewness and sample kurtosis.

In almost all countries the distribution of income is skewed to the right.

Other measures of skewness

Pearson's skewness coefficients

Karl Pearson suggested simpler calculations as a measure of skewness:^[4] the Pearson mode or first skewness coefficient, defined by

- $(\text{mean} - \text{mode}) / \text{standard deviation}$,
- as well as Pearson's median or second skewness coefficient, defined by
- $3(\text{mean} - \text{median}) / \text{standard deviation}$.

The latter is a simple multiple of the nonparametric skew.

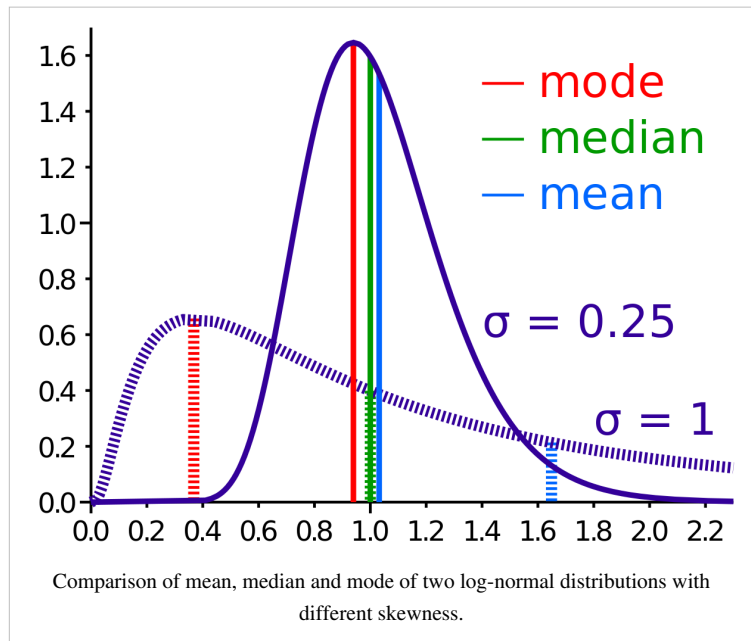
Starting from a standard cumulant expansion around a Normal distribution, one can actually show that skewness = $6(\text{mean} - \text{median}) / \text{standard deviation} (1 + \text{kurtosis} / 8) + O(\text{skewness}^2)$.^[citation needed] One should keep in mind that above given

equalities often don't hold even approximately and these empirical formulas are abandoned nowadays. There is no guarantee that these will be the same sign as each other or as the ordinary definition of skewness.

The adjusted Fisher-Pearson standardized moment coefficient is the version found in Excel and several statistical packages including Minitab, SAS and SPSS.^[5] The formula for this statistic is

$$G = \frac{n}{(n - 1)(n - 2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

where n is the sample size and s is the sample standard deviation.



Quantile based measures

A skewness function

$$\gamma(u) = \frac{F^{-1}(u) + F^{-1}(1-u) - 2F^{-1}(1/2)}{F^{-1}(u) - F^{-1}(1-u)}$$

can be defined,^{[6][7]} where F is the cumulative distribution function. This leads to a corresponding overall measure of skewness^[6] defined as the supremum of this over the range $1/2 \leq u < 1$. Another measure can be obtained by integrating the numerator and denominator of this expression. The function $\gamma(u)$ satisfies $-1 \leq \gamma(u) \leq 1$ and is well defined without requiring the existence of any moments of the distribution.

Galton's measure of skewness^[8] is $\gamma(u)$ evaluated at $u = 3/4$. Other names for this same quantity are the Bowley Skewness,^[9] the Yule-Kendall index^[10] and the quartile skewness.

Kelley's measure of skewness uses $u = 0.1$.^[citation needed]

L-moments

Use of L-moments in place of moments provides a measure of skewness known as the L-skewness.

Cyhelský's skewness coefficient

An alternative skewness coefficient may be derived from the sample mean and the individual observations:

$$a = (\text{number of observations below the mean} - \text{number of observations above the mean}) / \text{total number of observations}$$

The distribution of the skewness coefficient a in large sample sizes (≥ 45) approaches that of a normal distribution. If the variates have a normal or a uniform distribution the distribution of a is the same. The behavior of a when the variates have other distributions is currently unknown. Although this measure of skewness is very intuitive, an analytic approach to its distribution has proven difficult.

Distance skewness

A value of skewness equal to zero does not imply that the probability distribution is symmetric. Thus there is a need for another measure of asymmetry which has this property: such a measure was introduced in 2000.^[11] It is called **distance skewness** and denoted by dSkew. If X is a random variable which takes values in the d -dimensional Euclidean space, X has finite expectation, X' is an independent identically distributed copy of X and $\| \cdot \|$ denotes the norm in the Euclidean space then a simple *measure of asymmetry* is

$$\text{dSkew}(X) := 1 - \mathbb{E}\|X - X'\| / \mathbb{E}\|X + X'\| \text{ if } X \text{ is not } 0 \text{ with probability one,}$$

and $\text{dSkew}(X) := 0$ for $X = 0$ (with probability 1). Distance skewness is always between 0 and 1, equals 0 if and only if X is diagonally symmetric (X and $-X$ has the same probability distribution) and equals 1 if and only if X is a nonzero constant with probability one.^[12] Thus there is a simple consistent statistical test of diagonal symmetry based on the **sample distance skewness**:

$$\text{dSkew}_n(X) := 1 - \sum_{i,j} \|x_i - x_j\| / \sum_{i,j} \|x_i + x_j\|.$$

Groeneveld & Meeden's coefficient

Groeneveld & Meeden have suggested, as an alternative measure of skewness,

$$\text{skew}(X) = \frac{(\mu - \nu)}{E(|X - \nu|)}$$

where μ is the mean, ν is the median, $|\dots|$ is the absolute value and $E()$ is the expectation operator.

Notes

- [1] Susan Dean, Barbara Illowsky "Descriptive Statistics: Skewness and the Mean, Median, and Mode" (<http://cnx.org/content/m17104/latest/>), Connexions website
- [2] Duncan Cramer (1997) *Fundamental Statistics for Social Research*. Routledge. ISBN13 9780415172042 (p 85)
- [3] Kendall, M.G.; Stuart, A. (1969) *The Advanced Theory of Statistics, Volume 1: Distribution Theory, 3rd Edition*, Griffin. ISBN10 0-85264-141-9 (Ex 12.9)
- [4] <http://www.stat.upd.edu.ph/s114%20notes%20fcapistrano/Chapter%2010.pdf>
- [5] Doane DP, Seward LE (2011) *J Stat Educ* 19 (2)
- [6] MacGillivray (1992)
- [7] Hinkley DV (1975) "On power transformations to symmetry", *Biometrika*, 62, 101–111
- [8] Johnson *et al* (1994) p3, p40
- [9] Kenney JF and Keeping ES (1962) *Mathematics of Statistics, Pt. 1, 3rd ed.*, Van Nostrand, (page 102)
- [10] Wilks DS (1995) *Statistical Methods in the Atmospheric Sciences*, p27. Academic Press. ISBN 0-12-751965-3
- [11] Szekely, G.J. (2000). "Pre-limit and post-limit theorems for statistics", In: *Statistics for the 21st Century* (eds. C. R. Rao and G. J. Szekely), Dekker, New York, pp. 411–422.
- [12] Szekely, G.J. and Mori, T.F. (2001) "A characteristic measure of asymmetry and its application for testing diagonal symmetry", *Communications in Statistics - Theory and Methods* 30/8&9, 1633–1639.

References

- Johnson, NL, Kotz, S, Balakrishnan N (1994) *Continuous Univariate Distributions, Vol 1, 2nd Edition* Wiley ISBN 0-471-58495-9
- MacGillivray, HL (1992). "Shape properties of the g- and h- and Johnson families". *Comm. Statistics — Theory and Methods* 21: 1244–1250.

External links

- Hazewinkel, Michiel, ed. (2001), "Asymmetry coefficient" (<http://www.encyclopediaofmath.org/index.php?title=p/a013590>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- An Asymmetry Coefficient for Multivariate Distributions (<http://petitjeanmichel.free.fr/itoweb.petitjean.skewness.html>) by Michel Petitjean
- On More Robust Estimation of Skewness and Kurtosis (<http://repositories.cdlib.org/cgi/viewcontent.cgi?article=1017&context=ucsdecon>) Comparison of skew estimators by Kim and White.
- Closed-skew Distributions — Simulation, Inversion and Parameter Estimation (<http://dahoiv.net/master/index.html>)

Kurtosis

In probability theory and statistics, **kurtosis** (from the Greek word *κυρτός*, *kyrtos* or *kurtos*, meaning curved, arching) is any measure of the "peakedness" of the probability distribution of a real-valued random variable.^[1] In a similar way to the concept of skewness, *kurtosis* is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population. There are various interpretations of kurtosis, and of how particular measures should be interpreted; these are primarily peakedness (width of peak), tail weight, and lack of shoulders (distribution primarily peak and tails, not in between).

One common measure of kurtosis, originating with Karl Pearson, is based on a scaled version of the fourth moment of the data or population, but it has been argued that this really measures heavy tails, and not peakedness.^[2] For this measure, higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations. It is common practice to use an adjusted version of Pearson's kurtosis, the **excess kurtosis**, to provide a comparison of the shape of a given distribution to that of the normal distribution. Distributions with negative or positive excess kurtosis are called **platykurtic distributions** or **leptokurtic distributions** respectively.

Alternative measures of kurtosis are: the L-kurtosis, which is a scaled version of the fourth L-moment; measures based on 4 population or sample quantiles.^[3] These correspond to the alternative measures of skewness that are not based on ordinary moments.

Pearson moments

The fourth standardized moment is defined as

$$\beta_2 = \frac{E[(X - \mu)^4]}{(E[(X - \mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

where μ_4 is the fourth moment about the mean and σ is the standard deviation. The fourth standardized moment is lower bounded by the squared skewness plus 1^[4]

$$\frac{\mu_4}{\sigma^4} \geq \left(\frac{\mu_3}{\sigma^3}\right)^2 + 1$$

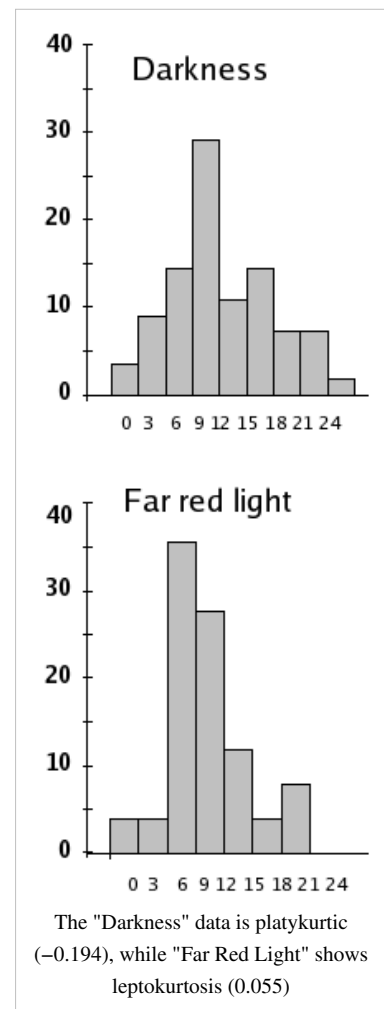
where μ_3 is the third moment about the mean.

The fourth standardized moment is sometimes used as the definition of kurtosis in older works, but is not the definition used here.

Kurtosis is more commonly defined as the fourth cumulant divided by the square of the second cumulant^[citation needed], which is equal to the fourth moment around the mean divided by the square of the variance of the probability distribution minus 3,

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3$$

which is also known as **excess kurtosis**. The "minus 3" at the end of this formula is often explained as a correction to make the kurtosis of the normal distribution equal to zero. Another reason can be seen by looking at the formula for the kurtosis of the sum of random variables. Suppose that Y is the sum of n identically distributed independent random variables all with the same distribution as X . Then



$$\text{Kurt}[Y] = \frac{\kappa_4(Y)}{\kappa_2(Y)^2} = \frac{n\kappa_4(X)}{(n\kappa_2(X))^2} = \frac{1}{n} \frac{\kappa_4(X)}{\kappa_2(X)^2} = \frac{1}{n} \text{Kurt}[X].$$

This formula would be much more complicated if kurtosis were defined just as μ_4 / σ^4 (without the minus 3).

More generally, if X_1, \dots, X_n are independent random variables, not necessarily identically distributed, but all having the same variance, then

$$\text{Kurt} \left(\sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{Kurt}(X_i),$$

whereas this identity would not hold if the definition did not include the subtraction of 3.

The fourth standardized moment must be at least 1, so the excess kurtosis must be -2 or more. This lower bound is realized by the Bernoulli distribution with $p = 1/2$, or "coin toss". There is no upper limit to the excess kurtosis and it may be infinite.

Interpretation

The exact interpretation of the Pearson measure of kurtosis (or excess kurtosis) is disputed. The "classical" interpretation, which applies only to symmetric and unimodal distributions (those whose skewness is 0), is that kurtosis measures both the "peakedness" of the distribution and the heaviness of its tail.^[5] Various statisticians have proposed other interpretations, such as "lack of shoulders" (where the "shoulder" is defined vaguely as the area between the peak and the tail, or more specifically as the area about one standard deviation from the mean) or "bimodality".^[6] Balanda and MacGillivray assert that the standard definition of kurtosis "is a poor measure of the kurtosis, peakedness, or tail weight of a distribution"^[7] and instead propose to "define kurtosis vaguely as the location- and scale-free movement of probability mass from the shoulders of a distribution into its center and tails".

Terminology and examples

A high kurtosis distribution has a sharper *peak* and longer, fatter *tails*, while a low kurtosis distribution has a more rounded peak and shorter, thinner tails.

Distributions with zero excess kurtosis are called **mesokurtic**, or mesokurtotic. The most prominent example of a mesokurtic distribution is the normal distribution family, regardless of the values of its parameters. A few other well-known distributions can be mesokurtic, depending on parameter values: for example the binomial distribution is mesokurtic for $p = 1/2 \pm \sqrt{1/12}$.

A distribution with positive excess kurtosis is called **leptokurtic**, or leptokurtotic. "Lepto-" means "slender".^[8] In terms of shape, a leptokurtic distribution has a more acute *peak* around the mean and *fatter tails*. Examples of leptokurtic distributions include the Student's t-distribution, Rayleigh distribution, Laplace distribution, exponential distribution, Poisson distribution and the logistic distribution. Such distributions are sometimes termed *super Gaussian*.^[citation needed]

A distribution with negative excess kurtosis is called **platykurtic**, or platykurtotic. "Platy-" means "broad".^[9] In terms of shape, a platykurtic distribution has a lower, wider *peak* around the mean and *thinner tails*. Examples of platykurtic distributions include the continuous or discrete uniform distributions, and the raised cosine distribution. The most platykurtic distribution of all is the Bernoulli distribution with $p = 1/2$ (for example the number of times one obtains "heads" when flipping a coin once, a coin toss), for which the excess kurtosis is -2 . Such distributions are sometimes termed *sub-Gaussian*.^[10]



The coin toss is the most platykurtic distribution

Graphical examples

The Pearson type VII family

The effects of kurtosis are illustrated using a parametric family of distributions whose kurtosis can be adjusted while their lower-order moments and cumulants remain constant. Consider the Pearson type VII family, which is a special case of the Pearson type IV family restricted to symmetric densities. The probability density function is given by

$$f(x; a, m) = \frac{\Gamma(m)}{a \sqrt{\pi} \Gamma(m - 1/2)} \left[1 + \left(\frac{x}{a}\right)^2 \right]^{-m},$$

where a is a scale parameter and m is a shape parameter.

All densities in this family are symmetric. The k th moment exists provided $m > (k + 1)/2$. For the kurtosis to exist, we require $m > 5/2$. Then the mean and skewness exist and are both identically zero. Setting $a^2 = 2m - 3$ makes the variance equal to unity. Then the only free parameter is m , which controls the fourth moment (and cumulant) and hence the kurtosis. One can reparameterize with $m = 5/2 + 3/\gamma_2$, where γ_2 is the kurtosis as defined above. This yields a one-parameter leptokurtic family with zero mean, unit variance, zero skewness, and arbitrary positive kurtosis. The reparameterized density is

$$g(x; \gamma_2) = f(x; a = \sqrt{2 + 6/\gamma_2}, m = 5/2 + 3/\gamma_2).$$

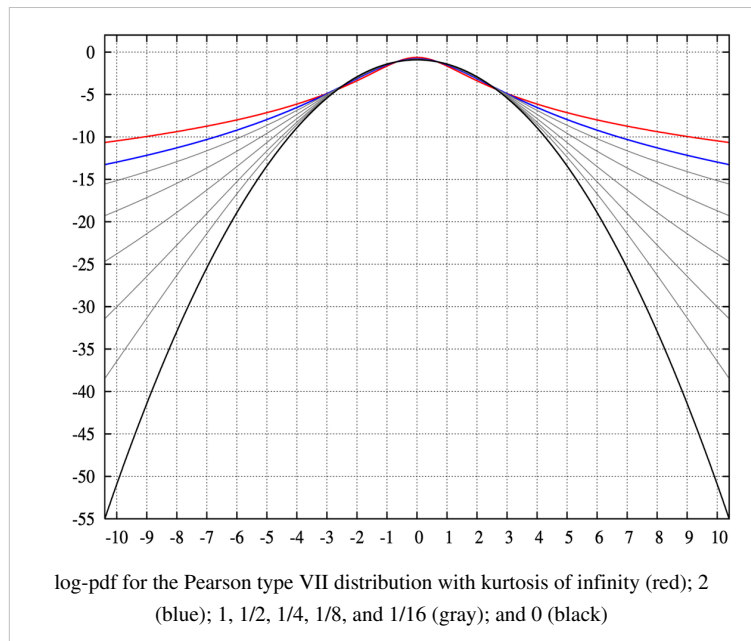
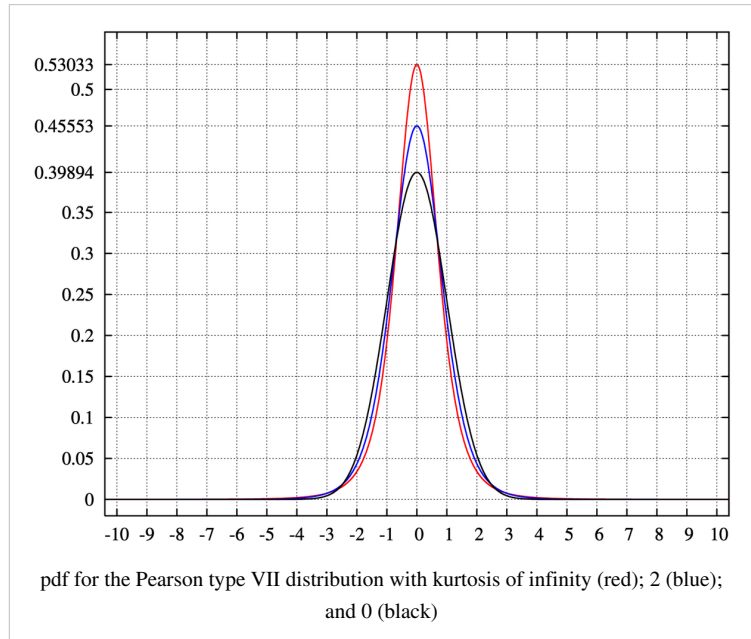
In the limit as $\gamma_2 \rightarrow \infty$ one obtains the density

$$g(x) = 3 \left(2 + x^2 \right)^{-5/2},$$

which is shown as the red curve in the images on the right.

In the other direction as $\gamma_2 \rightarrow 0$ one obtains the standard normal density as the limiting distribution, shown as the black curve.

In the images on the right, the blue curve represents the density $x \mapsto g(x; 2)$ with kurtosis of 2. The top image shows that leptokurtic densities in this family have a higher peak than the mesokurtic normal density. The comparatively fatter tails of the leptokurtic densities are illustrated in the second image, which plots the natural logarithm of the Pearson type VII densities: the black curve is the logarithm of the standard normal density, which is a parabola. One can see that the normal density allocates little probability mass to the regions far from the mean ("has thin tails"), compared with the blue curve of the leptokurtic Pearson type VII density with kurtosis of 2.

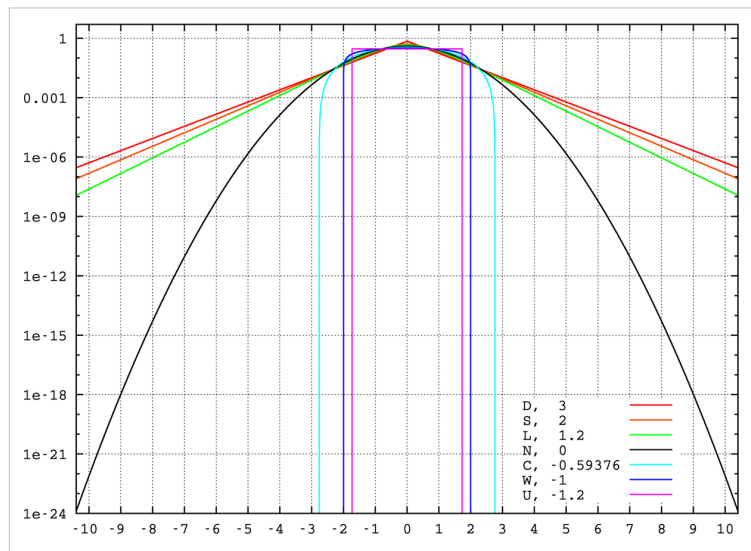
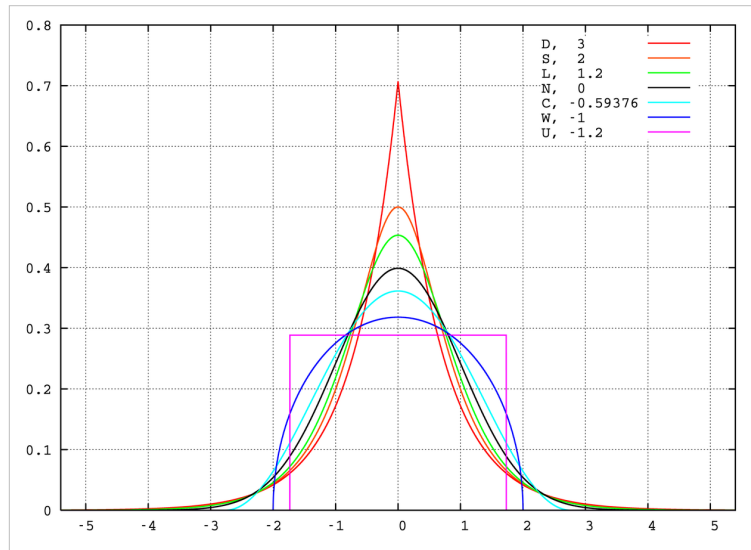


Between the blue curve and the black are other Pearson type VII densities with $\gamma_2 = 1, 1/2, 1/4, 1/8,$ and $1/16$. The red curve again shows the upper limit of the Pearson type VII family, with $\gamma_2 = \infty$ (which, strictly speaking, means that the fourth moment does not exist). The red curve decreases the slowest as one moves outward from the origin ("has fat tails").

Kurtosis of well-known distributions

Several well-known, unimodal and symmetric distributions from different parametric families are compared here. Each has a mean and skewness of zero. The parameters have been chosen to result in a variance equal to 1 in each case. The images on the right show curves for the following seven densities, on a linear scale and logarithmic scale:

- D: Laplace distribution, also known as the double exponential distribution, red curve (two straight lines in the log-scale plot), excess kurtosis = 3
- S: hyperbolic secant distribution, orange curve, excess kurtosis = 2
- L: logistic distribution, green curve, excess kurtosis = 1.2
- N: normal distribution, black curve (inverted parabola in the log-scale plot), excess kurtosis = 0
- C: raised cosine distribution, cyan curve, excess kurtosis = $-0.593762\dots$
- W: Wigner semicircle distribution, blue curve, excess kurtosis = -1
- U: uniform distribution, magenta curve (shown for clarity as a rectangle in both images), excess kurtosis = -1.2 .



Note that in these cases the platykurtic densities have bounded support, whereas the densities with positive or zero excess kurtosis are supported on the whole real line.

There exist platykurtic densities with infinite support,

- e.g., exponential power distributions with sufficiently large shape parameter b

and there exist leptokurtic densities with finite support.

- e.g., a distribution that is uniform between -3 and -0.3 , between -0.3 and 0.3 , and between 0.3 and 3 , with the same density in the $(-3, -0.3)$ and $(0.3, 3)$ intervals, but with 20 times more density in the $(-0.3, 0.3)$ interval

Sample kurtosis

For a sample of n values the **sample excess kurtosis** is

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3$$

where m_4 is the fourth sample moment about the mean, m_2 is the second sample moment about the mean (that is, the sample variance), x_i is the i^{th} value, and \bar{x} is the sample mean.

The variance of the sample kurtosis of a sample of size n from the normal distribution is^[11]

$$\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}$$

An approximate alternative is $24/n$ but this is inaccurate for small samples.

Estimators of population kurtosis

Given a sub-set of samples from a population, the sample excess kurtosis above is a biased estimator of the population excess kurtosis. The usual estimator of the population excess kurtosis (used in DAP/SAS, Minitab, PSPP/SPSS, and Excel but not by BMDP) is G_2 , defined as follows:

$$\begin{aligned} G_2 &= \frac{k_4}{k_2^2} \\ &= \frac{n^2 \left((n+1) m_4 - 3(n-1) m_2^2 \right) (n-1)^2}{(n-1)(n-2)(n-3) n^2 m_2^2} \\ &= \frac{n-1}{(n-2)(n-3)} \left((n+1) \frac{m_4}{m_2^2} - 3(n-1) \right) \\ &= \frac{n-1}{(n-2)(n-3)} \left((n+1) g_2 + 6 \right) \\ &= \frac{(n+1)n(n-1)}{(n-2)(n-3)} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3 \frac{(n-1)^2}{(n-2)(n-3)} \\ &= \frac{(n+1)n}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{k_2^2} - 3 \frac{(n-1)^2}{(n-2)(n-3)} \end{aligned}$$

where k_4 is the unique symmetric unbiased estimator of the fourth cumulant, k_2 is the unbiased estimate of the second cumulant (identical to the unbiased estimate of the sample variance), m_4 is the fourth sample moment about the mean, m_2 is the second sample moment about the mean, x_i is the i^{th} value, and \bar{x} is the sample mean. Unfortunately, G_2 is itself generally biased. For the normal distribution it is unbiased.^[citation needed]

For computationally efficient ways of calculating the sample kurtosis see Algorithms for calculating higher-order statistics.

Applications

D'Agostino's K-squared test is a goodness-of-fit normality test based on a combination of the sample skewness and sample kurtosis, as is the Jarque–Bera test for normality.

For non-normal samples, the variance of the variance depends on the kurtosis; for details, please see variance.

Pearson's definition of kurtosis is used as an indicator of intermittency in turbulence.^[12]

Other measures of kurtosis

A different measure of "kurtosis", that is of the "peakedness" of a distribution, is provided by using L-moments instead of the ordinary moments.

References

- [1] Dodge, Y. (2003) *The Oxford Dictionary of Statistical Terms*. OUP. ISBN 0-19-920613-9
- [2] SAS Elementary Statistics Procedures (<http://support.sas.com/onlinedoc/913/getDoc/en/proc.hlp/a002473332.htm>), SAS Institute (section on Kurtosis)
- [3] Joanes & Gill (1998)
- [4] Petitjean M. (2013), "The Chiral Index: Applications to Multivariate Distributions and to 3D molecular graphs", Proceedings of 12th International Symposium on Operational Research in Slovenia SOR'13, pp. 11-16, L. Zadnik Stirn, J. Zerovnik, J. Povh, S. Drobne, A. Lisec, Eds., Slovenian Society INFORMATIKA (SDI), Section for Operations Research (SOR), ISBN 978-961-6165-40-2
- [5] Balanda, Kevin P. and H.L. MacGillivray (1988), "Kurtosis: A Critical Review", *The American Statistician*, 42:2, pp. 111–119.
- [6] Darlington, Richard B. (1970), "Is Kurtosis Really 'Peakedness'?", *The American Statistician*, 24:2, pp. 19–22.
- [7] Balanda and MacGillivray, p. 114.
- [8] <http://medical-dictionary.thefreedictionary.com/lepto->
- [9] <http://www.yourdictionary.com/platy-prefix>
- [10] The original paper presenting sub-Gaussians J.P. Kahane, "Local properties of functions in terms of random Fourier series," *Stud. Math.*, 19, No. i, 1-25 (1960). See also Buldygin, V. V., & Kozachenko, Y. V. (1980). "Sub-Gaussian random variables". *Ukrainian Mathematical Journal*, 32(6), 483-489.
- [11] Duncan Cramer (1997) *Fundamental Statistics for Social Research*. Routledge. ISBN13: 9780415172042 (p 89)
- [12] Sandborn 1958 <http://dx.doi.org/10.1017/S0022112059000581>

Further reading

- Joanes, D. N. & Gill, C. A. (1998) Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society (Series D): The Statistician* **47** (1), 183–189. doi: 10.1111/1467-9884.00122 (<http://dx.doi.org/10.1111/1467-9884.00122>)
- Kim, Tae-Hwan; & White, Halbert. (2003/4). "On More Robust Estimation of Skewness and Kurtosis: Simulation and Application to the S&P500 Index". (<http://escholarship.org/uc/item/7b52v07p>) *Finance Research Letters*, 1, 56–70 doi: 10.1016/S1544-6123(03)00003-5 ([http://dx.doi.org/10.1016/S1544-6123\(03\)00003-5](http://dx.doi.org/10.1016/S1544-6123(03)00003-5)) Alternative source (http://weber.ucsd.edu/~hwhite/pub_files/hwcv-092.pdf) (Comparison of kurtosis estimators)
- Seier, E. & Bonett, D.G. (2003). Two families of kurtosis measures. *Metrika*, 58, 59–70.

External links

- Hazewinkel, Michiel, ed. (2001), "Excess coefficient" (<http://www.encyclopediaofmath.org/index.php?title=p/e036800>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Free Online Software (Calculator) (<http://www.wessa.net/skewkurt.wasp>) computes various types of skewness and kurtosis statistics for any dataset (includes small and large sample tests)..
- Kurtosis (<http://jeff560.tripod.com/k.html>) on the Earliest known uses of some of the words of mathematics (<http://jeff560.tripod.com/mathword.html>)
- Celebrating 100 years of Kurtosis (<http://faculty.etsu.edu/seier/doc/Kurtosis100years.doc>) a history of the topic, with different measures of kurtosis.

Ranking

A **ranking** is a relationship between a set of items such that, for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second. In mathematics, this is known as a weak order or total preorder of objects. It is not necessarily a total order of objects because two different objects can have the same ranking. The rankings themselves are totally ordered. For example, materials are totally preordered by hardness, while degrees of hardness are totally ordered.

By reducing detailed measures to a sequence of ordinal numbers, rankings make it possible to evaluate complex information according to certain criteria. Thus, for example, an Internet search engine may rank the pages it finds according to an estimation of their relevance, making it possible for the user quickly to select the pages they are likely to want to see.

Analysis of data obtained by ranking commonly requires non-parametric statistics.

Strategies for assigning rankings

It is not always possible to assign rankings uniquely. For example, in a race or competition two (or more) entrants might tie for a place in the ranking. When computing an ordinal measurement, two (or more) of the quantities being ranked might measure equal. In these cases, one of the strategies shown below for assigning the rankings may be adopted.

A common shorthand way to distinguish these ranking strategies is by the ranking numbers that would be produced for four items, with the first item ranked ahead of the second and third (which compare equal) which are both ranked ahead of the fourth. These names are also shown below.

Standard competition ranking ("1224" ranking)

In competition ranking, items that compare equal receive the same ranking number, and then a gap is left in the ranking numbers. The number of ranking numbers that are left out in this gap is one less than the number of items that compared equal. Equivalently, each item's ranking number is 1 plus the number of items ranked above it. This ranking strategy is frequently adopted for competitions, as it means that if two (or more) competitors tie for a position in the ranking, the position of all those ranked below them is unaffected (i.e., a competitor only comes second if exactly one person scores better than them, third if exactly two people score better than them, fourth if exactly three people score better than them, etc.).

Thus if A ranks ahead of B and C (which compare equal) which are both ranked ahead of D, then A gets ranking number 1 ("first"), B gets ranking number 2 ("joint second"), C also gets ranking number 2 ("joint second") and D gets ranking number 4 ("fourth").

Modified competition ranking ("1334" ranking)

Sometimes, competition ranking is done by leaving the gaps in the ranking numbers *before* the sets of equal-ranking items (rather than after them as in standard competition ranking). The number of ranking numbers that are left out in this gap remains one less than the number of items that compared equal. Equivalently, each item's ranking number is equal to the number of items ranked equal to it or above it. This ranking ensures that a competitor only comes second if they score higher than all but one of their opponents, third if they score higher than all but two of their opponents, etc.

Thus if A ranks ahead of B and C (which compare equal) which are both ranked ahead of D, then A gets ranking number 1 ("first"), B gets ranking number 3 ("joint third"), C also gets ranking number 3 ("joint third") and D gets ranking number 4 ("fourth"). In this case, nobody would get ranking number 2 ("second") and that would be left as a gap.

Dense ranking ("1223" ranking)

In dense ranking, items that compare equal receive the same ranking number, and the next item(s) receive the immediately following ranking number. Equivalently, each item's ranking number is 1 plus the number of items ranked above it that are distinct with respect to the ranking order.

Thus if A ranks ahead of B and C (which compare equal) which are both ranked ahead of D, then A gets ranking number 1 ("first"), B gets ranking number 2 ("joint second"), C also gets ranking number 2 ("joint second") and D gets ranking number 3 ("third").

Ordinal ranking ("1234" ranking)

In ordinal ranking, all items receive distinct ordinal numbers, including items that compare equal. The assignment of distinct ordinal numbers to items that compare equal can be done at random, or arbitrarily, but it is generally preferable to use a system that is arbitrary but consistent, as this gives stable results if the ranking is done multiple times. An example of an arbitrary but consistent system would be to incorporate other attributes into the ranking order (such as alphabetical ordering of the competitor's name) to ensure that no two items exactly match.

With this strategy, if A ranks ahead of B and C (which compare equal) which are both ranked ahead of D, then A gets ranking number 1 ("first") and D gets ranking number 4 ("fourth"), and **either** B gets ranking number 2 ("second") and C gets ranking number 3 ("third") **or** C gets ranking number 2 ("second") and B gets ranking number 3 ("third").

In computer data processing, ordinal ranking is also referred to as "row numbering"....

Fractional ranking ("1 2.5 2.5 4" ranking)

Items that compare equal receive the same ranking number, which is the mean of what they would have under ordinal rankings. Equivalently, the ranking number of 1 plus the number of items ranked above it plus half the number of items equal to it. This strategy has the property that the sum of the ranking numbers is the same as under ordinal ranking. For this reason, it is used in computing Borda counts and in statistical tests (see below).

Thus if A ranks ahead of B and C (which compare equal) which are both ranked ahead of D, then A gets ranking number 1 ("first"), B and C each get ranking number 2.5 (average of "joint second/third") and D gets ranking number 4 ("fourth").

Here's an example: Suppose you have the data set 1 1 2 3 3 4 5 5 There are 5 different numbers, so there would be five different ranks. If 1 and 1 were actually different numbers, they would occupy ranks 1 and 2. Since they are the same number, you find their rank by finding the average as follows : (rank) 1 + (rank) 2 / 2 numbers total = 1.5 (average rank). The next number in the data set, 2, is thus assigned the rank of 3 (the average takes up 1 and 2 in the first two 1's). The two 3's in the set would occupy ranks 3 and 4 if they were different numbers, so the average rank

would be computed as follows: $(4 + 5) / 2 = 4.5$. 4 would get the rank of 6 (because your average took into account rank 4 and 5 in the average). there are 3 5's in the data set. Their average rank is computed as $(7+8+9)/3 = 8$

Your ranks would be: 1.5 1.5 3 4.5 4.5 6 8 8 8

Ranking in statistics

In statistics, "ranking" refers to the data transformation in which numerical or ordinal values are replaced by their rank when the data are sorted. For example, the numerical data 3.4, 5.1, 2.6, 7.3 are observed, the ranks of these data items would be 2, 3, 1 and 4 respectively. For example, the ordinal data hot, cold, warm would be replaced by 3, 1, 2. In these examples, the ranks are assigned to values in ascending order. (In some other cases, descending ranks are used.) Ranks are related to the indexed list of order statistics, which consists of the original dataset rearranged into ascending order.

Some kinds of statistical tests employ calculations based on ranks. Examples include:

- Friedman test
- Kruskal-Wallis test
- Rank products
- Spearman's rank correlation coefficient
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test

Some ranks can have non-integer values for tied data values. For example, when there is an even number of copies of the same data value, the above described fractional statistical rank of the tied data ends in $\frac{1}{2}$.

Rank function in Excel

The **rank** function in Microsoft Excel assigns competition ranks ("1224") as described above. For some statistical purposes, that is not the desired result - for instance, it means that the sum of ranks for a list of a given length changes depending on the number of ties. Pottel has described a user defined ranking function which assigns fractional ranks to ties to keep the sum consistent.^[1]

Examples of ranking

- In politics, rankings focus on the comparison of economic, social, environmental and governance performance of countries, see List of international rankings
- In many sports, individuals or teams are given rankings, generally by the sport's governing body
 - In football (soccer) national teams are ranked in the FIFA World Rankings and, unofficially, in the World Football Elo Ratings.
 - In the Olympic Games, each member country (NOC) is ranked based upon gold, silver and bronze medal counts in the Olympic medal rankings.
 - In snooker, players are ranked using the Snooker world rankings
 - In ice hockey, national teams are ranked in the IIHF World Ranking
 - In golf, the top male golfers are ranked using the Official World Golf Rankings
- In relation to credit standing, the ranking of a security refers to where that particular security would stand in a wind up of the issuing company, i.e., its seniority in the company's capital structure. For instance, capital notes are subordinated securities; they would rank behind senior debt in a wind up. In other words the holders of senior debt would be paid out before subordinated debt holders received any funds.
- Search engines rank web pages by their expected relevance to a user's query using a combination of query-dependent and query-independent methods. Query-independent methods attempt to measure the estimated importance of a page, independent of any consideration of how well it matches the specific query.

Query-independent ranking is usually based on link analysis; examples include the HITS algorithm, PageRank and TrustRank. Query-dependent methods attempt to measure the degree to which a page matches a specific query, independent of the importance of the page. Query-dependent ranking is usually based on heuristics that consider the number and locations of matches of the various query words on the page itself, in the URL or in any anchor text referring to the page.

- In Webometrics it is possible to rank institutions according to their presence in the web (number of webpages) and the impact of these contents (external inlinks=site citations), such as the Webometrics Ranking of World Universities
- In video gaming, players may be given a ranking. To "rank up" is to achieve a higher ranking relative to other players, especially with strategies that do not depend on the player's skill.
- The TrueSkill ranking system is a skill based ranking system for Xbox Live developed at Microsoft Research
- A bibliogram ranks common noun phrases in a piece of text.
- In language, the status of an item (usually through what is known as "downranking" or "rank-shifting") in relation to the uppermost rank in a clause; for example, in the sentence "I want to eat the cake you made today", "eat" is on the uppermost rank, but "made" is downranked as part of the nominal group "the cake you made today"; this nominal group behaves as though it were a single noun (i.e., I want to eat *it*), and thus the verb within it ("made") is ranked differently from "eat".
- Academic journals are sometimes ranked according to impact factor; the number of later articles that cite articles in a given journal.

References

- [1] Hans Pottel. Statistical flaws in Excel (<http://www.mis.coventry.ac.uk/~nhunt/pottel.pdf>)

External links

- Ronen Perry, The Relative Value of American Law Reviews: A Critical Appraisal of Ranking Methods (<http://papers.ssrn.com/abstract=806144>)
- Ronen Perry, The Relative Value of American Law Reviews: Refinement and Implementation (<http://papers.ssrn.com/abstract=897063>)
- A MATLAB Toolbox for computing rankings using five different methodologies (<http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=19496>)
- TrueSkill Ranking System (<http://research.microsoft.com/en-us/projects/trueskill/default.aspx>)
- Ranking Library written in Ruby (<http://github.com/quidproquo/ranker>)

Graphics

Box plot

In descriptive statistics, a **box plot** or **boxplot** is a convenient way of graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (*whiskers*) indicating variability outside the upper and lower quartiles, hence the terms **box-and-whisker plot** and **box-and-whisker diagram**. Outliers may be plotted as individual points.

Box plots display differences between populations without making any assumptions of the underlying statistical distribution: they are non-parametric. The spacings between the different parts of the box help indicate the degree of dispersion (spread) and skewness in the data, and identify outliers. In addition to the points themselves, they allow one to visually estimate various L-estimators, notably the interquartile range, midhinge, range, mid-range, and trimean. Boxplots can be drawn either horizontally or vertically.

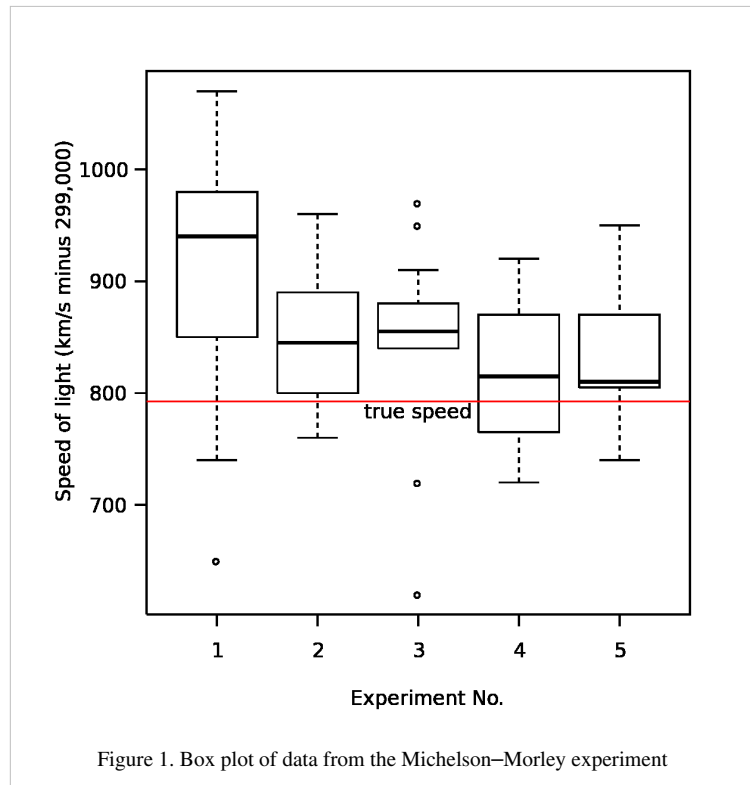


Figure 1. Box plot of data from the Michelson–Morley experiment

Types of boxplots

Box and whisker plots are uniform in their use of the box: the bottom and top of the box are always the first and third quartiles, and the band inside the box is always the second quartile (the median). But the ends of the whiskers can represent several possible alternative values, among them:

- the minimum and maximum of all of the data (as in Figure 2)
- the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile (as in Figure 3)
- one standard deviation above and below the mean of the data
- the 9th percentile and the 91st percentile
- the 2nd percentile and the 98th percentile.

Any data not included between the whiskers should be plotted as an outlier with a dot, small circle, or star, but occasionally this is not done.

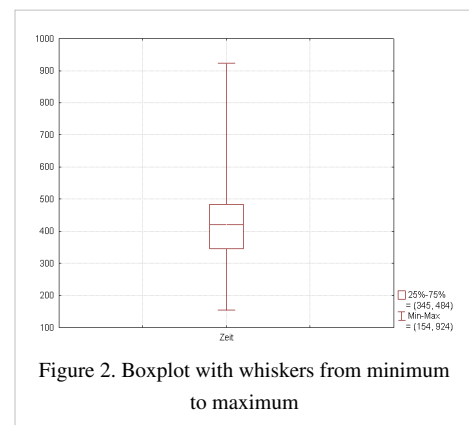


Figure 2. Boxplot with whiskers from minimum to maximum

Some box plots include an additional character to represent the mean of the data.

On some box plots a crosshatch is placed on each whisker, before the end of the whisker.

Rarely, box plots can be presented with no whiskers at all.

Because of this variability, it is appropriate to describe the convention being used for the whiskers and outliers in the caption for the plot.

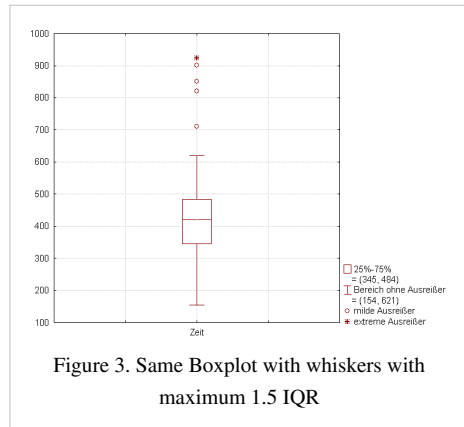


Figure 3. Same Boxplot with whiskers with maximum 1.5 IQR

The unusual percentiles 2%, 9%, 91%, 98% are sometimes used for whisker cross-hatches and whisker ends to show the seven-number summary. If the data is normally distributed, the locations of the seven marks on the box plot will be equally spaced.

Variations

Since the American mathematician John W. Tukey introduced this type of visual data display in 1969, several variations on the traditional box plot have been described. Two of the most common are variable width box plots and notched box plots (see figure 4).

Variable width box plots illustrate the size of each group whose data is being plotted by making the width of the box proportional to the size of the group. A popular convention is to make the box width proportional to the square root of the size of the group.

Notched box plots apply a "notch" or narrowing of the box around the median. Notches are useful in offering a rough guide to significance of difference of medians; if the notches of two boxes do not overlap, this offers evidence of a statistically significant difference between the medians. The width of the notches is proportional to the interquartile range of the sample and inversely proportional to the square root of the size of the sample. However, there is uncertainty about the most appropriate multiplier (as this may vary depending on the similarity of the variances of the samples). One convention is to use $\pm 1.58 \times IQR \div \sqrt{n}$.

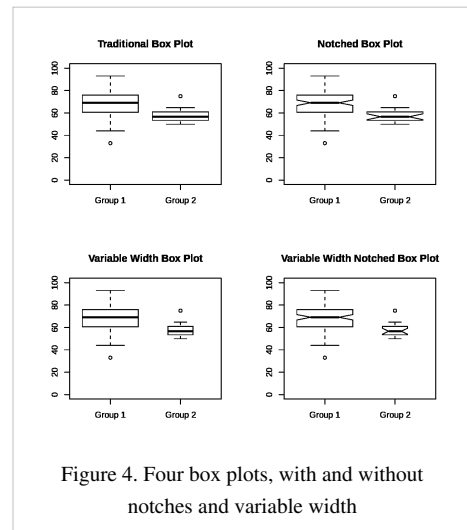


Figure 4. Four box plots, with and without notches and variable width

Visualization

The box plot is a quick way of examining one or more sets of data graphically. Box plots may seem more primitive than a histogram or kernel density estimate but they do have some advantages. They take up less space and are therefore particularly useful for comparing distributions between several groups or sets of data (see Figure 1 for an example). Choice of number and width of bins techniques can heavily influence the appearance of a histogram, and choice of bandwidth can heavily influence the appearance of a kernel density estimate.

As looking at a statistical distribution is more intuitive than looking at a box plot, comparing the box plot against the probability density function (theoretical histogram) for a normal $N(0,1\sigma^2)$ distribution may be a useful tool for understanding the box plot (Figure 5).

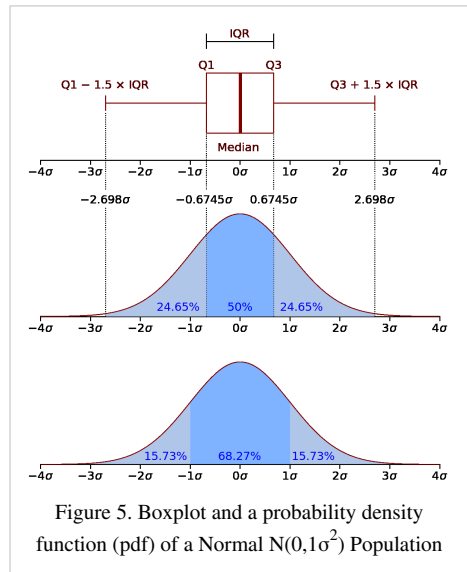


Figure 5. Boxplot and a probability density function (pdf) of a Normal $N(0,1\sigma^2)$ Population

References

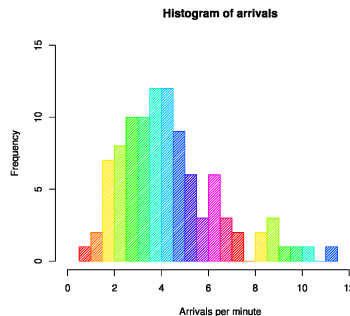
Further reading

- John W. Tukey (1977). *Exploratory Data Analysis*. Addison-Wesley.
- Benjamini, Y. (1988). "Opening the Box of a Boxplot". *The American Statistician* **42** (4): 257–262. doi: 10.2307/2685133 (<http://dx.doi.org/10.2307/2685133>). JSTOR 2685133 (<http://www.jstor.org/stable/2685133>).
- Rousseeuw, P. J.; Ruts, I.; Tukey, J. W. (1999). "The Bagplot: A Bivariate Boxplot". *The American Statistician* **53** (4): 382–387. doi: 10.2307/2686061 (<http://dx.doi.org/10.2307/2686061>). JSTOR 2686061 (<http://www.jstor.org/stable/2686061>).

External links

- Visual Presentation of Data by Means of Box Plots (<http://www.lcgceurope.com/lcgceurope/data/articlestandard/lcgceurope/132005/152912/article.pdf>)
- On-line box plot calculator with explanations and examples (<http://www.physics.csbsju.edu/stats/box2.html>) (Has beeswarm example)
- Beeswarm Boxplot (<http://www.r-statistics.com/2011/03/beeswarm-boxplot-and-plotting-it-with-r/>) - superimposing a frequency-jittered stripchart on top of a boxplot

Histogram

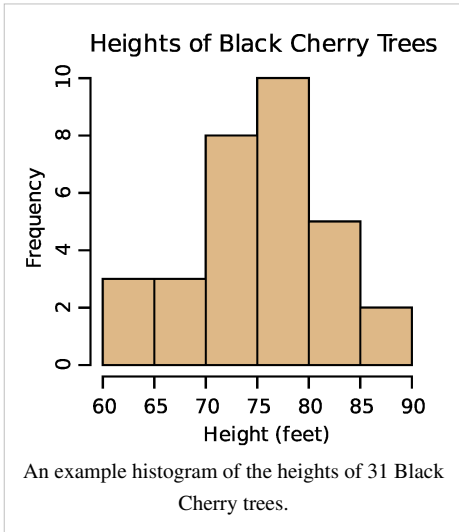
Histogram																											
 <p style="text-align: center;">Histogram of arrivals</p> <p>The histogram shows the frequency of arrivals per minute. The x-axis is labeled 'Arrivals per minute' and ranges from 0 to 12. The y-axis is labeled 'Frequency' and ranges from 0 to 15. The bars are colored in a rainbow sequence. The highest frequency is 12 for 4 arrivals per minute.</p> <table border="1" style="display: none;"> <caption>Data for Histogram of arrivals</caption> <thead> <tr> <th>Arrivals per minute</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>2</td></tr> <tr><td>3</td><td>7</td></tr> <tr><td>4</td><td>10</td></tr> <tr><td>5</td><td>12</td></tr> <tr><td>6</td><td>9</td></tr> <tr><td>7</td><td>6</td></tr> <tr><td>8</td><td>3</td></tr> <tr><td>9</td><td>2</td></tr> <tr><td>10</td><td>3</td></tr> <tr><td>11</td><td>1</td></tr> <tr><td>12</td><td>1</td></tr> </tbody> </table>		Arrivals per minute	Frequency	1	1	2	2	3	7	4	10	5	12	6	9	7	6	8	3	9	2	10	3	11	1	12	1
Arrivals per minute	Frequency																										
1	1																										
2	2																										
3	7																										
4	10																										
5	12																										
6	9																										
7	6																										
8	3																										
9	2																										
10	3																										
11	1																										
12	1																										
First described by	Karl Pearson																										
Purpose	To roughly assess the probability distribution of a given variable by depicting the frequencies of observations occurring in certain ranges of values																										

In statistics, a **histogram** is a graphical representation of the distribution of data. It is an estimate of the probability distribution of a continuous variable and was first introduced by Karl Pearson. A histogram is a representation of tabulated frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval. The height of a rectangle is also equal to the frequency density of the interval, i.e., the frequency divided by the width of the interval. The total area of the histogram is equal to the number of data. A histogram may also be normalized displaying relative frequencies. It then shows the proportion of cases that fall into each of several categories, with the total area equaling 1. The categories are usually specified as consecutive, non-overlapping intervals of a variable. The categories (intervals) must be adjacent, and often are chosen to be of the same size.^[1] The rectangles of a histogram are drawn so that they touch each other to indicate that the original variable is continuous.^[2]

Histograms are used to plot the density of data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the x -axis are all 1, then a histogram is identical to a relative frequency plot.

An alternative to the histogram is kernel density estimation, which uses a kernel to smooth samples. This will construct a smooth probability density function, which **will in general more accurately reflect the underlying variable**. The histogram is one of the seven basic tools of quality control.

Etymology

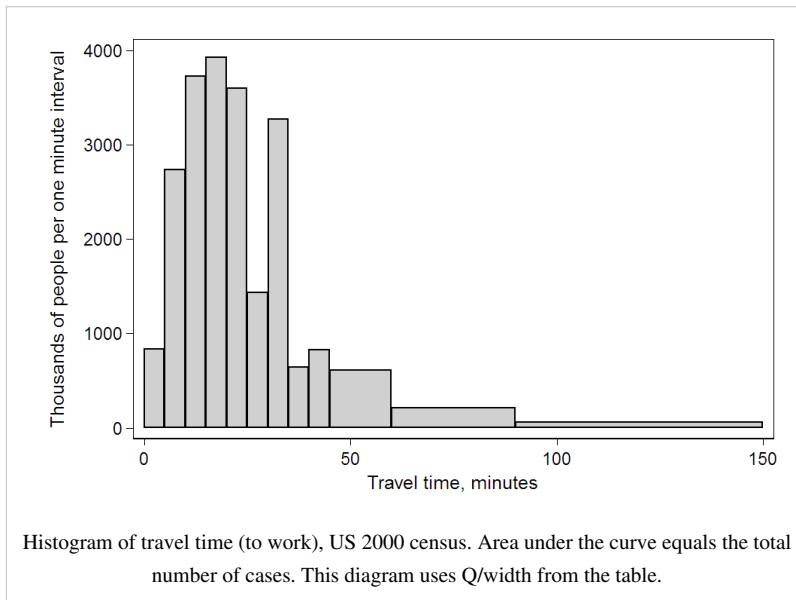


The etymology of the word *histogram* is uncertain. Sometimes it is said to be derived from the Greek *histos* 'anything set upright' (as the masts of a ship, the bar of a loom, or the vertical bars of a histogram); and *gramma* 'drawing, record, writing'. It is also said that Karl Pearson, who introduced the term in 1891, derived the name from "historical diagram".

Examples

The U.S. Census Bureau found that there were 124 million people who work outside of their homes.^[3] Using their data on the time occupied by travel to work, Table 2 below shows the absolute number of people who responded with travel times "at least 30 but less than 35 minutes" is higher than the numbers for the categories above and below it. This

is likely due to people rounding their reported journey time.^[citation needed] The problem of reporting values as somewhat arbitrarily rounded numbers is a common phenomenon when collecting data from people.^[citation needed]

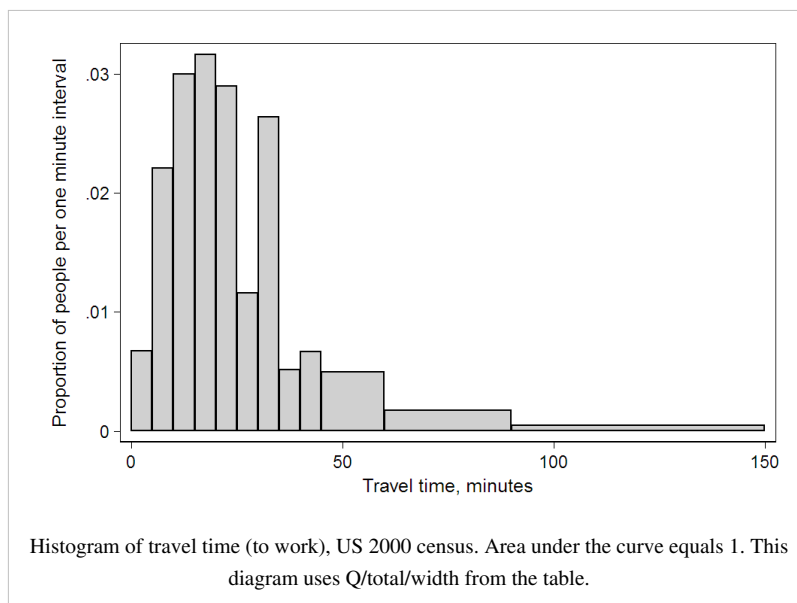


Data by absolute numbers

Interval	Width	Quantity	Quantity/width
0	5	4180	836
5	5	13687	2737
10	5	18618	3723
15	5	19634	3926
20	5	17981	3596
25	5	7190	1438
30	5	16369	3273
35	5	3212	642

40	5	4122	824
45	15	9200	613
60	30	6461	215
90	60	3435	57

This histogram shows the number of cases per unit interval as the height of each block, so that the area of each block is equal to the number of people in the survey who fall into its category. The area under the curve represents the total number of cases (124 million). This type of histogram shows absolute numbers, with Q in thousands.



Data by proportion

Interval	Width	Quantity (Q)	$Q/\text{total}/\text{width}$
0	5	4180	0.0067
5	5	13687	0.0221
10	5	18618	0.0300
15	5	19634	0.0316
20	5	17981	0.0290
25	5	7190	0.0116
30	5	16369	0.0264
35	5	3212	0.0052
40	5	4122	0.0066
45	15	9200	0.0049
60	30	6461	0.0017
90	60	3435	0.0005

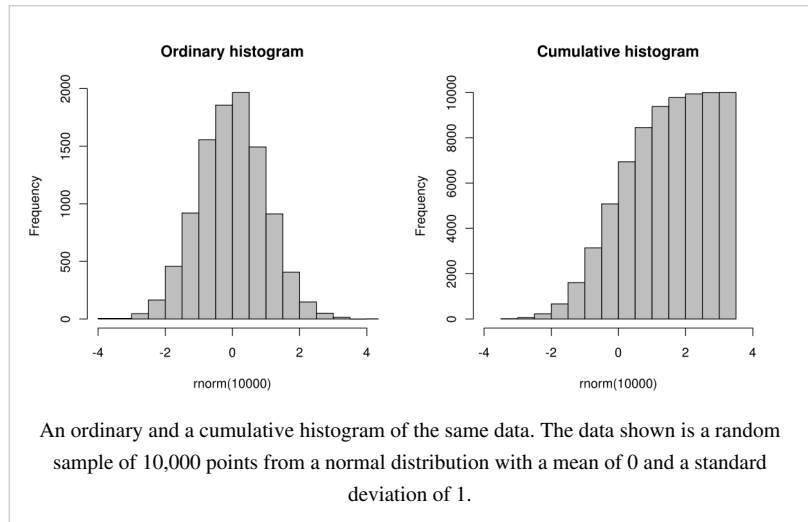
This histogram differs from the first only in the vertical scale. The area of each block is the fraction of the total that each category represents, and the total area of all the bars is equal to 1 (the fraction meaning "all"). The curve displayed is a simple density estimate. This version shows proportions, and is also known as a unit area histogram.

In other words, a histogram represents a frequency distribution by means of rectangles whose widths represent class intervals and whose areas are proportional to the corresponding frequencies: the height of each is the average frequency density for the interval. The intervals are placed together in order to show that the data represented by the histogram, while exclusive, is also contiguous. (E.g., in a histogram it is possible to have two connecting intervals of 10.5–20.5 and 20.5–33.5, but not two connecting intervals of 10.5–20.5 and 22.5–32.5. Empty intervals are represented as empty and not skipped.)^[4]

Mathematical definition

In a more general mathematical sense, a histogram is a function m_i that counts the number of observations that fall into each of the disjoint categories (known as *bins*), whereas the graph of a histogram is merely one way to represent a histogram. Thus, if we let n be the total number of observations and k be the total number of bins, the histogram m_i meets the following conditions:

$$n = \sum_{i=1}^k m_i.$$



Cumulative histogram

A cumulative histogram is a mapping that counts the cumulative number of observations in all of the bins up to the specified bin. That is, the cumulative histogram M_i of a histogram m_j is defined as:

$$M_i = \sum_{j=1}^i m_j.$$

Number of bins and width

There is no "best" number of bins, and different bin sizes can reveal different features of the data. Grouping data is at least as old as Graunt's work in the 17th century, but no systematic guidelines were given until Sturges's work in 1926.

Using wider bins where the density is low reduces noise due to sampling randomness; using narrower bins where the density is high (so the signal drowns the noise) gives greater precision to the density estimation. Thus varying the bin-width within a histogram can be beneficial. Nonetheless, equal-width bins are widely used.

Some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution. Depending on the actual data distribution and the goals of the analysis, different bin widths may be appropriate, so experimentation is usually needed to determine an appropriate width. There are, however, various useful guidelines and rules of thumb.^[5]

The number of bins k can be assigned directly or can be calculated from a suggested bin width h as:

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil.$$

The braces indicate the ceiling function.

Square-root choice

$$k = \sqrt{n},$$

which takes the square root of the number of data points in the sample (used by Excel histograms and many others).^[6]

Sturges' formula

Sturges' formula is derived from a binomial distribution and implicitly assumes an approximately normal distribution.

$$k = \lceil \log_2 n + 1 \rceil,$$

It implicitly bases the bin sizes on the range of the data and can perform poorly if $n < 30$.^[citation needed] It may also perform poorly if the data are not normally distributed.

Rice Rule

$$k = \lceil 2n^{1/3} \rceil,$$

The Rice Rule^[7] is presented as a simple alternative to Sturges's rule.

Doane's formula

Doane's formula^[8] is a modification of Sturges' formula which attempts to improve its performance with non-normal data.

$$k = 1 + \log_2(n) + \log_2 \left(1 + \frac{|g_1|}{\sigma_{g_1}} \right)$$

where g_1 is the estimated 3rd-moment-skewness of the distribution and

$$\sigma_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$$

Scott's normal reference rule

$$h = \frac{3.5\hat{\sigma}}{n^{1/3}},$$

where $\hat{\sigma}$ is the sample standard deviation. Scott's normal reference rule is optimal for random samples of normally distributed data, in the sense that it minimizes the integrated mean squared error of the density estimate.

Freedman–Diaconis' choice

The Freedman–Diaconis rule is:

$$h = 2 \frac{\text{IQR}(x)}{n^{1/3}},$$

which is based on the interquartile range, denoted by IQR. It replaces 3.5σ of Scott's rule with 2IQR , which is less sensitive than the standard deviation to outliers in data.

Choice based on minimization of an estimated L^2 risk function

$$\arg \min_h \frac{2\bar{m} - v}{h^2}$$

where \bar{m} and v are mean and biased variance of a histogram with bin-width h , $\bar{m} = \frac{1}{k} \sum_{i=1}^k m_i$ and $v = \frac{1}{k} \sum_{i=1}^k (m_i - \bar{m})^2$.

Remark

A good reason why the number of bins should be proportional to $n^{1/3}$ is the following: suppose that the data are obtained as n independent realizations of a bounded probability distribution with smooth density. Then the histogram remains equally »rugged« as n tends to infinity. If s is the »width« of the distribution (e. g., the standard deviation or the inter-quartile range), then the number of units in a bin (the frequency) is of order nh/s and the

relative standard error is of order $\sqrt{s/(nh)}$. Comparing to the next bin, the relative change of the frequency is of order h provided that the derivative of the density is non-zero. These two are of the same order if h is of order $s/n^{1/3}$, so that k is of order $n^{1/3}$.

This simple cubic root choice can also be applied to bins with non-constant width.

References

- [1] Howitt, D. and Cramer, D. (2008) *Statistics in Psychology*. Prentice Hall
- [2] Charles Stangor (2011) "Research Methods For The Behavioral Sciences". Wadsworth, Cengage Learning. ISBN 9780840031976.
- [3] US 2000 census (<http://www.census.gov/prod/2004pubs/c2kbr-33.pdf>).
- [4] Dean, S., & Illowsky, B. (2009, February 19). Descriptive Statistics: Histogram. Retrieved from the Connexions Web site: <http://cnx.org/content/m16298/1.11/>
- [5] e.g. § 5.6 "Density Estimation", W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S* (2002), Springer, 4th edition. ISBN 0-387-95457-0.
- [6] EXCEL 2007: Histogram (<http://cameron.econ.ucdavis.edu/excel/ex11/histogram.html>)
- [7] Online Statistics Education: A Multimedia Course of Study (<http://onlinestatbook.com/>). Project Leader: David M. Lane, Rice University (chapter 2 "Graphing Distributions", section "Histograms")
- [8] Doane DP (1976) Aesthetic frequency classification. *American Statistician*, 30: 181–183

Further reading

- Lancaster, H.O. *An Introduction to Medical Statistics*. John Wiley and Sons. 1974. ISBN 0-471-51250-8

External links

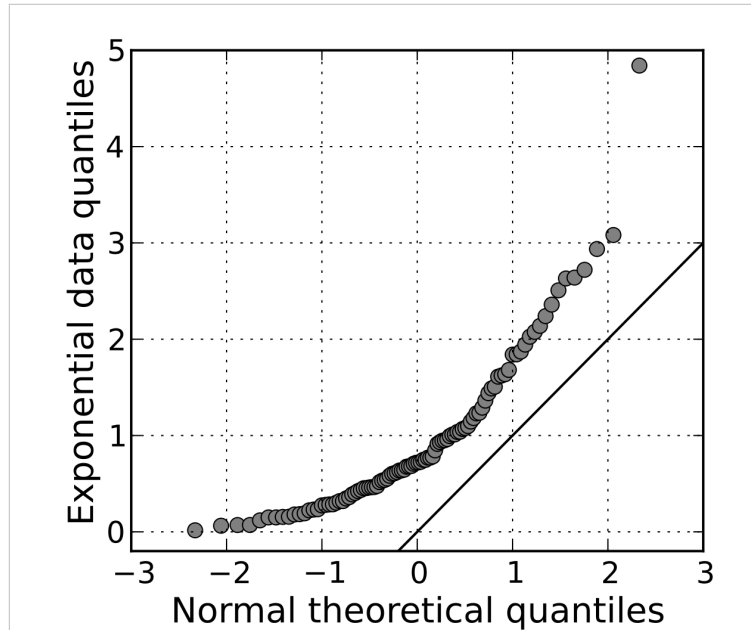
- Journey To Work and Place Of Work (<http://www.census.gov/population/www/socdemo/journey.html>) (*location of census document cited in example*)
- Smooth histogram for signals and images from a few samples (<http://www.mathworks.com/matlabcentral/fileexchange/30480-histconnect>)
- Histograms: Construction, Analysis and Understanding with external links and an application to particle Physics. (<http://quarknet.fnal.gov/toolkits/ati/histograms.html>)
- A Method for Selecting the Bin Size of a Histogram (<http://2000.jukuin.keio.ac.jp/shimazaki/res/histogram.html>)
- Interactive histogram generator (<http://www.shodor.org/interactivate/activities/histogram/>)
- Matlab function to plot nice histograms (<http://www.mathworks.com/matlabcentral/fileexchange/27388-plot-and-compare-nice-histograms-by-default>)
- Dynamic Histogram in MS Excel (<http://excelandfinance.com/histogram-in-excel/>)
- Histogram construction (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_ModelerActivities_MixtureModel_1) and manipulation (http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_PowerTransformFamily_Graphs) using Java applets, and charts (http://www.socr.ucla.edu/htmls/SOCR_Charts.html) on SOCR

Q-Q plot

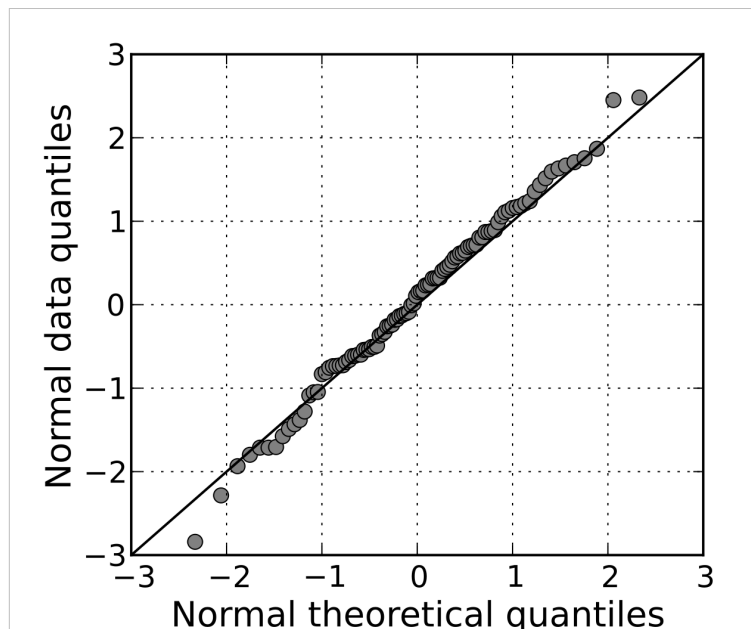
In statistics, a **Q-Q plot** ("Q" stands for *quantile*) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (x,y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the (number of the) interval for the quantile.

If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q-Q plots can be used to compare collections of data, or theoretical distributions. The use of Q-Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q-Q plot is generally a more powerful approach to doing this than the common technique of comparing histograms of the two samples, but requires more skill to interpret. Q-Q plots are commonly used to compare a data set to a theoretical model.^[1] This can provide an assessment of "goodness of fit" that is graphical, rather than reducing to a numerical summary. Q-Q plots are also used to compare two theoretical



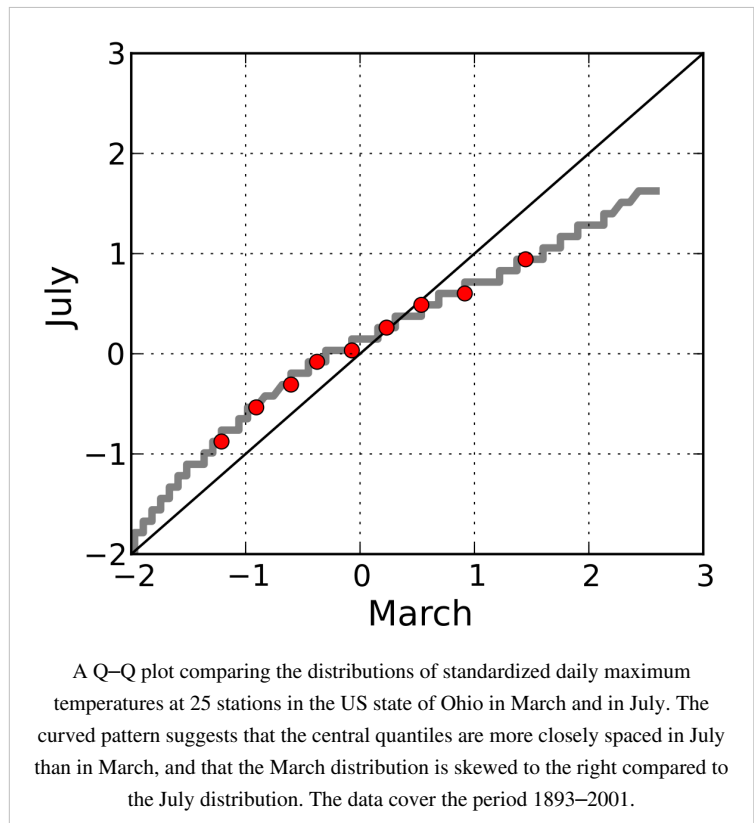
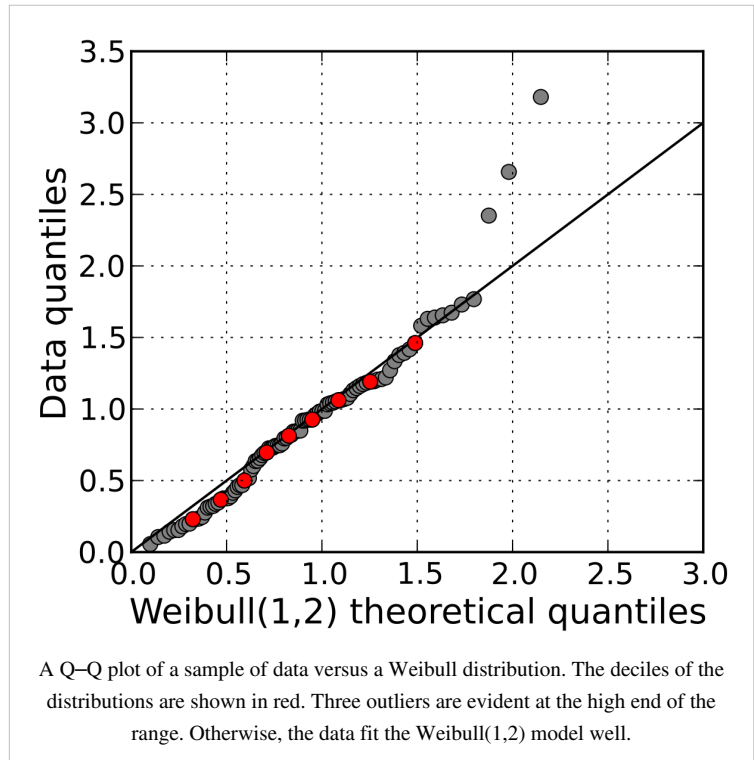
A normal Q-Q plot of randomly generated, independent standard exponential data, ($X \sim \text{Exp}(1)$). This Q-Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis. The points follow a strongly nonlinear pattern, suggesting that the data are not distributed as a standard normal ($X \sim N(0,1)$). The offset between the line and the points suggests that the mean of the data is not 0. The median of the points can be determined to be near 0.7



A normal Q-Q plot comparing randomly generated, independent standard normal data on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points suggests that the data are normally distributed.

distributions to each other. Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.

The term "probability plot" sometimes refers specifically to a Q–Q plot, sometimes to a more general class of plots, and sometimes to the less commonly used P–P plot. The **probability plot correlation coefficient** is a quantity derived from the idea of Q–Q plots, which measures the agreement of a fitted distribution with observed data and which is sometimes used as a means of fitting a distribution to data.



Definition and construction

A **Q–Q plot** is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted. If one or both of the axes in a Q–Q plot is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are uniquely defined and can be obtained by inverting the CDF. If a theoretical probability distribution with a discontinuous CDF is one of the two distributions being compared, some of the quantiles may not be defined, so an interpolated quantile may be plotted. If the Q–Q plot is based on data, there are multiple quantile estimators in use. Rules for forming Q–Q plots when quantiles must be estimated or interpolated are called plotting positions.

A simple case is where one has two data sets of the same size. In that case, to make the Q–Q plot, one orders each set in increasing order, then pairs off and plots the corresponding values. A more complicated construction is the case where two data sets of different sizes are being compared. To construct the Q–Q plot in this case, it is necessary to use an interpolated quantile estimate so that quantiles corresponding to the same underlying probability can be constructed.

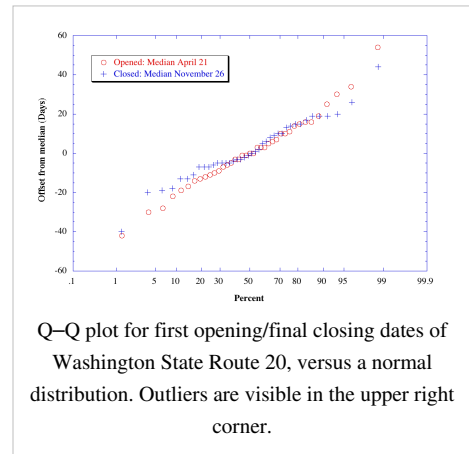
More abstractly, given two cumulative probability distribution functions F and G , with associated quantile functions F^{-1} and G^{-1} (the inverse function of the CDF is the quantile function), the Q–Q plot draws the q th quantile of F against the q th quantile of G for a range of values of q . Thus, the Q–Q plot is a parametric curve indexed over $[0,1]$ with values in the real plane \mathbf{R}^2 .

Interpretation

The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right. If the two distributions being compared are identical, the Q–Q plot follows the 45° line $y = x$. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line $y = x$. If the general trend of the Q–Q plot is flatter than the line $y = x$, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis. Conversely, if the general trend of the Q–Q plot is steeper than the line $y = x$, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis. Q–Q plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.

Although a Q–Q plot is based on quantiles, in a standard Q–Q plot it is not possible to determine which point in the Q–Q plot determines a given quantile. For example, it is not possible to determine the median of either of the two distributions being compared by inspecting the Q–Q plot. Some Q–Q plots indicate the deciles to make determinations such as this possible.

The slope and position of a linear regression between the quantiles gives a measure of the relative location and relative scale of the samples. If the median of the distribution plotted on the horizontal axis is 0, the intercept of a regression line is a measure of location, and the slope is a measure of scale. The distance between medians is another measure of relative location reflected in a Q–Q plot. The "probability plot correlation coefficient" is the correlation coefficient between the paired sample quantiles. The closer the correlation coefficient is to one, the closer the distributions are to being shifted, scaled versions of each other. For distributions with a single shape parameter, the probability plot correlation coefficient plot (PPCC plot) provides a method for estimating the shape parameter – one simply computes the correlation coefficient for different values of the shape parameter, and uses the one with the



best fit, just as if one were comparing distributions of different types.

Another common use of Q–Q plots is to compare the distribution of a sample to a theoretical distribution, such as the standard normal distribution $N(0,1)$, as in a normal probability plot. As in the case when comparing two samples of data, one orders the data (formally, computes the order statistics), then plots them against certain quantiles of the theoretical distribution.

Plotting positions

The choice of quantiles from a theoretical distribution has occasioned much discussion. A natural choice, given a sample of size n , is k/n for $k = 1, \dots, n$, as these are the quantiles that the sampling distribution realizes. Unfortunately, the last of these, n/n , corresponds to the 100th percentile – the maximum value of the theoretical distribution, which is often infinite. To fix this, one may shift these over, using $(k - 0.5)/n$, or instead space the points evenly in the uniform distribution, using $k/(n + 1)$. This last one was suggested early on by Weibull, and recently it has been argued to be the definitive position by Lasse Makkonen. The claimed unique status of this estimator was rebutted by N.J. Cook.

For plotting positions, context matters. They are used for estimates of exceedance probabilities and other things as well, and there are disputes about whether the Weibull plotting position is the right procedure for all uses. Many other choices have been suggested, both formal and heuristic, based on theory or simulations relevant in context. The following subsections discuss some of these.

Expected value of the order statistic

In using a normal probability plot, the quantiles one uses are the rankits, the quantile of the expected value of the order statistic of a standard normal distribution.

More generally, Shapiro–Wilk test uses the expected values of the order statistics of the given distribution; the resulting plot and line yields the generalized least squares estimate for location and scale (from the intercept and slope of the fitted line).^[1] Although this is not too important for the normal distribution (the location and scale are estimated by the mean and standard deviation, respectively), it can be useful for many other distributions.

However, this requires calculating the expected values of the order statistic, which may be difficult if the distribution is not normal.

Median of the order statistics

Alternatively, one may use estimates of the *median* of the order statistics, which one can compute based on estimates of the median of the order statistics of a uniform distribution and the quantile function of the distribution; this was suggested by (Filliben 1975).

This can be easily generated for any distribution for which the quantile function can be computed, but conversely the resulting estimates of location and scale are no longer precisely the least squares estimates, though these only differ significantly for n small.

Heuristics

For the quantiles of the comparison distribution typically the formula $k/(n + 1)$ is used. Several different formulas have been used or proposed as symmetrical **plotting positions**. Such formulas have the form $(k - a)/(n + 1 - 2a)$ for some value of a in the range from 0 to 1/2, which gives a range between $k/(n + 1)$ and $(k - 1/2)/n$.

Other expressions include:

- $(k - 0.3) / (n + 0.4)$.
- $(k - 0.3175) / (n + 0.365)$.^[2]
- $(k - 0.326) / (n + 0.348)$.^[3]
- $(k - 1/3) / (n + 1/3)$.^[4]
- $(k - 0.375) / (n + 0.25)$.^[5]
- $(k - 0.4) / (n + 0.2)$.
- $(k - 0.44) / (n + 0.12)$.^[6]
- $(k - 0.567) / (n - 0.134)$.
- $(k - 1) / (n - 1)$.^[7]

For large sample size, n , there is little difference between these various expressions.

Filliben's estimate

The order statistic medians are the medians of the order statistics of the distribution. These can be expressed in terms of the quantile function and the order statistic medians for the continuous uniform distribution by:

$$N(i) = G(U(i))$$

where $U(i)$ are the uniform order statistic medians and G is the quantile function for the desired distribution. The quantile function is the inverse of the cumulative distribution function (probability that X is less than or equal to some value). That is, given a probability, we want the corresponding quantile of the cumulative distribution function.

James J. Filliben (Filliben 1975) uses the following estimates for the uniform order statistic medians:


$$m(i) = \begin{cases} 1 - m(n) & i = 1 \\ \frac{i - 0.3175}{n + 0.365} & i = 2, 3, \dots, n - 1 \\ 0.5^{1/n} & i = n. \end{cases}$$

The reason for this estimate is that the order statistic medians do not have a simple form.

Notes

- [1] Gnanadesikan (1977) p199.
- [2] Engineering Statistics Handbook: *Normal Probability Plot* (<http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm>) – Note that this also uses a different expression for the first & last points. (<http://engineering.tufts.edu/cee/people/vogel/publications/probability1986.pdf>) cites the original work by . This expression is an estimate of the medians of $U_{(k)}$.
- [3] *Distribution free plotting position*, Yu & Huang (<http://cat.inist.fr/?aModele=afficheN&cpsid=14151257>)
- [4] A simple (and easy to remember) formula for plotting positions; used in BMDP statistical package.
- [5] This is 's earlier approximation and is the expression used in MINITAB.
- [6] This plotting position was used by Irving I. Gringorten () to plot points in tests for the Gumbel distribution.
- [7] Used by , these plotting points are equal to the modes of $U_{(k)}$.

References

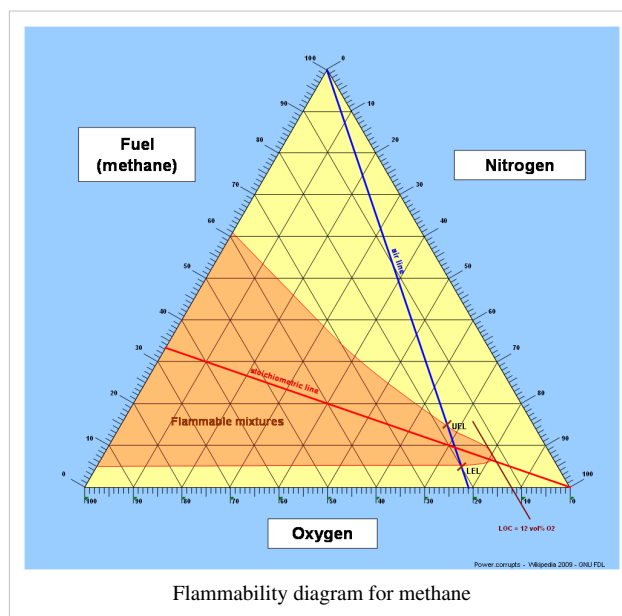
-  This article incorporates public domain material from websites or documents of the National Institute of Standards and Technology.
- Blom, G. (1958), *Statistical estimates and transformed beta variables*, New York: John Wiley and Sons
- Chambers, John; William Cleveland, Beat Kleiner, and Paul Tukey (1983), *Graphical methods for data analysis*, Wadsworth
- Cleveland, W.S. (1994) *The Elements of Graphing Data*, Hobart Press ISBN 0-9634884-1-4
- Filliben, J. J. (February 1975), "The Probability Plot Correlation Coefficient Test for Normality", *Technometrics* (American Society for Quality) **17** (1): 111–117, doi: 10.2307/1268008 (<http://dx.doi.org/10.2307/1268008>), JSTOR 1268008 (<http://www.jstor.org/stable/1268008>).
- Gibbons, Jean Dickinson; Chakraborti, Subhabrata (2003), *Nonparametric statistical inference* (<http://books.google.com/?id=kJbVO2G6VicC>) (4th ed.), CRC Press, ISBN 978-0-8247-4052-8
- Gnanadesikan, R. (1977) *Methods for Statistical Analysis of Multivariate Observations*, Wiley ISBN 0-471-30845-5.
- Thode, Henry C. (2002), *Testing for normality* (<http://books.google.com/?id=gbegXB4SdosC>), New York: Marcel Dekker, ISBN 0-8247-9613-6

External links

- Probability plot (<http://www.itl.nist.gov/div898/handbook/eda/section3/probplot.htm>)
- Alternate description of the QQ-Plot: http://www.stats.gla.ac.uk/steps/glossary/probability_distributions.html#qqplot

Ternary plot

A **ternary plot**, **ternary graph**, **triangle plot**, **simplex plot**, or **de Finetti diagram** is a barycentric plot on three variables which sum to a constant. It graphically depicts the ratios of the three variables as positions in an equilateral triangle. It is used in physical chemistry, petrology, mineralogy, metallurgy, and other physical sciences to show the compositions of systems composed of three species. In population genetics, it is often called a Gibbs triangle or a de Finetti diagram. In game theory, it is often called a *simplex plot*.^[*citation needed*]



In a ternary plot, the proportions of the three variables a , b , and c must sum to some constant, K . Usually, this constant is represented as 1.0 or 100%. Because $a + b + c = K$ for all substances being graphed, any one variable is not independent of the others, so only two variables must be known to find a sample's point on the graph: for instance, c must be equal to $K - a - b$. Because the three proportions cannot vary independently - there are only two degrees of freedom - it is possible to graph the intersection of all three variables in only two dimensions. ^[citation needed]

Reading values on the ternary plot

The advantage of using a ternary plot for depicting compositions is that three variables can be conveniently plotted in a two-dimensional graph. Ternary plots can also be used to create phase diagrams by outlining the composition regions on the plot where different phases exist.

Every point on a ternary plot represents a different composition of the three components. There are three common methods used to determine the ratios of the three species in the composition. The first method is an estimation based upon the phase diagram grid. The concentration of each species is 100% (pure phase) in its corner of the triangle and 0% at the line opposite it. The percentage of a specific species decreases linearly with increasing distance from this corner, as seen in figures 3–8. By drawing parallel lines at regular intervals between the zero line and the corner (as seen in the images), fine divisions can be established for easy estimation of the content of a species. For a given point, the fraction of each of the three materials in the composition can be determined by the first.

For phase diagrams that do not possess grid lines, the easiest way to determine the composition is to set the altitude of the triangle to 100% and determine the shortest distances from the point of interest to each of the three sides. The distances (the ratios of the distances to the total height of 100%) give the content of each of the species, as shown in figure 1.

The third method is based upon a larger number of measurements, but does not require the drawing of perpendicular lines. Straight lines are drawn from each corner, through the point of interest, to the corresponding side of the triangle. The lengths of these lines, as well as the lengths of the segments between the point and the corresponding sides, are measured individually. Ratios can then be determined by dividing these segments by the entire corresponding line as shown in the figure 2. (The sum of the ratios should add to 1).

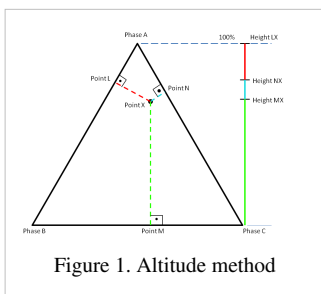
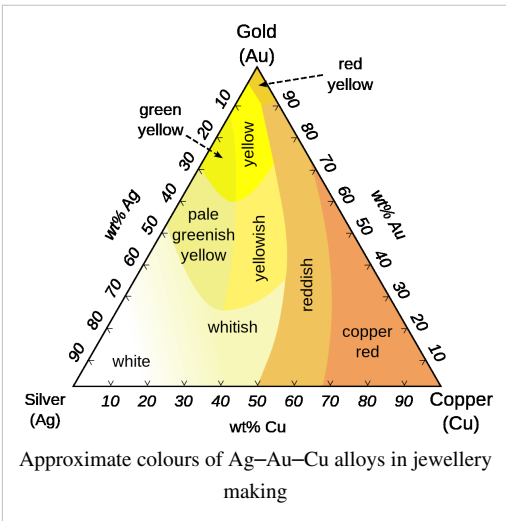


Figure 1. Altitude method

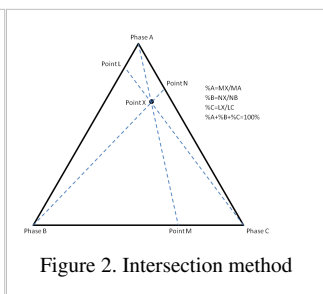


Figure 2. Intersection method

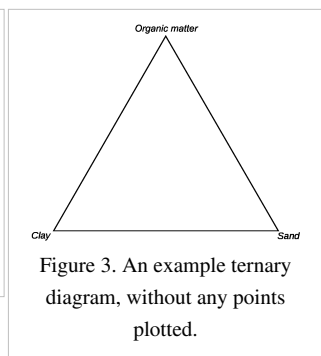


Figure 3. An example ternary diagram, without any points plotted.

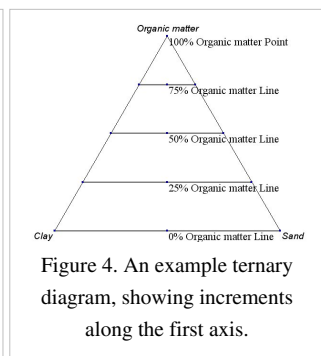


Figure 4. An example ternary diagram, showing increments along the first axis.

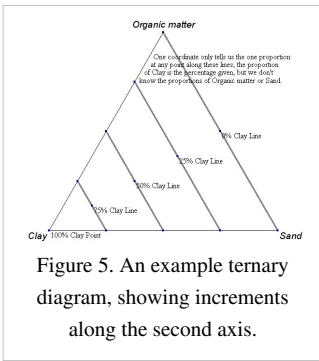


Figure 5. An example ternary diagram, showing increments along the second axis.

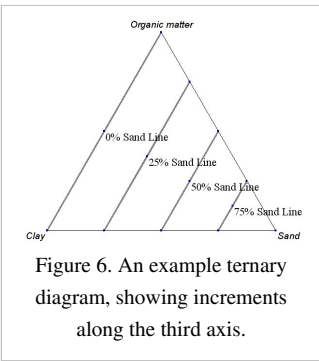


Figure 6. An example ternary diagram, showing increments along the third axis.

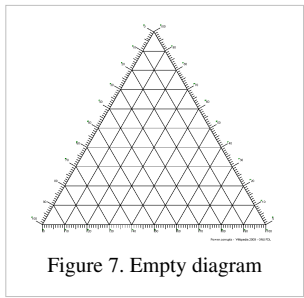


Figure 7. Empty diagram

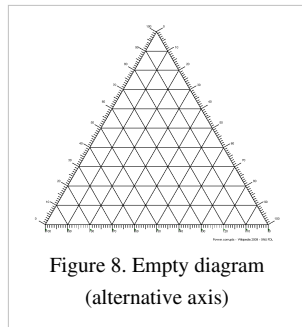
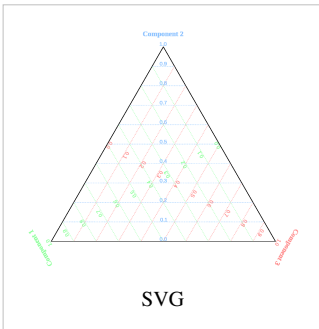


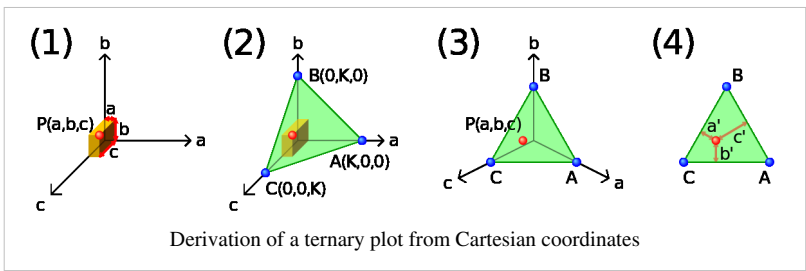
Figure 8. Empty diagram (alternative axis)



Derivation from Cartesian coordinates

Figure (1) shows an oblique projection of point $P(a,b,c)$ in a 3-dimensional Cartesian space with axes a, b and c , respectively.

If $a + b + c = K$ (a positive constant), P is restricted to a plane containing $A(K,0,0)$, $B(0,K,0)$ and $C(0,0,K)$. If a, b and c each cannot be negative, P is restricted to the triangle bounded by A, B and C , as in (2).



Derivation of a ternary plot from Cartesian coordinates

In (3), the axes are rotated to give an isometric view. The triangle, viewed face-on, appears equilateral.

In (4), the distances of P from lines BC, AC and AB are denoted by a', b' and c' , respectively.

For any line $\mathbf{l} = \mathbf{s} + t \hat{\mathbf{n}}$ in vector form ($\hat{\mathbf{n}}$ is a unit vector) and a point \mathbf{p} , the perpendicular distance from \mathbf{p} to \mathbf{l} is $\|(\mathbf{s} - \mathbf{p}) - ((\mathbf{s} - \mathbf{p}) \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}\|$.

In this case, point P is at $\mathbf{p} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$.

$$\text{Line BC has } \mathbf{s} = \begin{pmatrix} 0 \\ K \\ 0 \end{pmatrix} \text{ and } \hat{\mathbf{n}} = \frac{\begin{pmatrix} 0 \\ K \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ K \end{pmatrix}}{\left\| \begin{pmatrix} 0 \\ K \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ K \end{pmatrix} \right\|} = \frac{\begin{pmatrix} 0 \\ K \\ -K \end{pmatrix}}{\sqrt{0^2 + K^2 + (-K)^2}} = \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}.$$

Using the perpendicular distance formula,

$$\begin{aligned}
 a' &= \left\| \begin{pmatrix} -a \\ K-b \\ -c \end{pmatrix} - \left(\begin{pmatrix} -a \\ K-b \\ -c \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \right) \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \right\| \\
 &= \left\| \begin{pmatrix} -a \\ K-b \\ -c \end{pmatrix} - \left(0 + \frac{K-b}{\sqrt{2}} + \frac{c}{\sqrt{2}} \right) \begin{pmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \right\| \\
 &= \left\| \begin{pmatrix} -a \\ K-b - \frac{K-b+c}{2} \\ -c + \frac{K-b+c}{2} \end{pmatrix} \right\| = \left\| \begin{pmatrix} -a \\ \frac{K-b-c}{2} \\ \frac{K-b-c}{2} \end{pmatrix} \right\| \\
 &= \sqrt{(-a)^2 + \left(\frac{K-b-c}{2}\right)^2 + \left(\frac{K-b-c}{2}\right)^2} = \sqrt{a^2 + \frac{(K-b-c)^2}{2}}
 \end{aligned}$$

Substituting $K = a + b + c$,

$$a' = \sqrt{a^2 + \frac{(a+b+c-b-c)^2}{2}} = \sqrt{a^2 + \frac{a^2}{2}} = a\sqrt{\frac{3}{2}}.$$

Similar calculation on lines AC and AB gives

$$b' = b\sqrt{\frac{3}{2}} \text{ and } c' = c\sqrt{\frac{3}{2}}.$$

This shows that the distance of the point from the respective lines is linearly proportional to the original values a , b and c .^[1]

Plotting a ternary plot

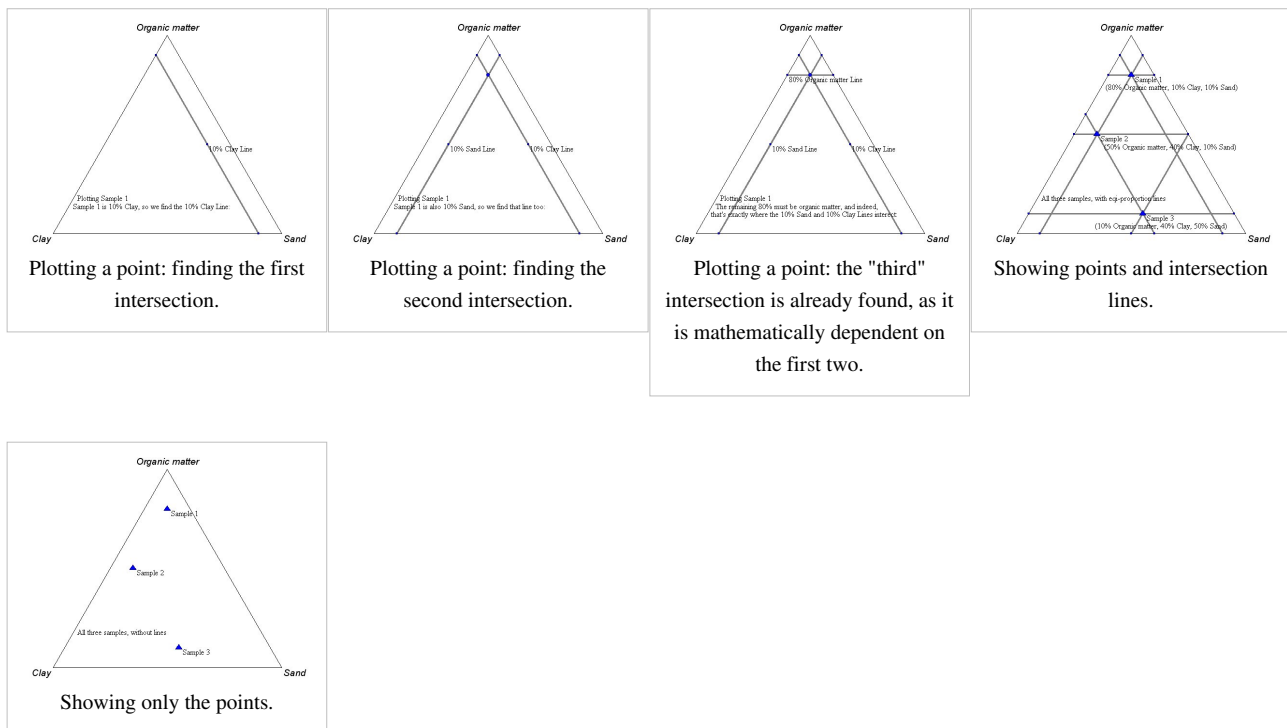
Cartesian coordinates are useful for plotting points in the triangle. Consider an equilateral ternary plot where $a = 100\%$ is placed at $(x, y) = (0, 0)$ and $b = 100\%$ at $(1, 0)$. Then $c = 100\%$ is $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$, and the triple (a, b, c) is $\left(\frac{1}{2} \frac{2b+c}{a+b+c}, \frac{\sqrt{3}}{2} \frac{c}{a+b+c}\right)$.

Example

This example shows how this works for a hypothetical set of three soil samples:

Sample #	Organic matter	Clay	Sand	Notes
Sample 1	80%	10%	10%	Because organic matter and clay make up 90% of this sample, the proportion of sand must be 10%.
Sample 2	50%	40%	10%	The proportion of sand is 10% in this sample too, but the proportions of organic matter and clay are different.
Sample 3	10%	40%	50%	This sample has the same proportion of clay as in Sample 2 does, but because it has a smaller proportion of organic matter, the proportion of sand must be larger, because all samples' proportions must sum to 100%.

Plotting the points



Software

Here is a list of software that help enable the creation of ternary plots

- JMP
- Origin
- R
- Veusz

References

Vaughan, Will (September 5, 2010). "Ternary plots"^[2]. Retrieved September 7, 2010.

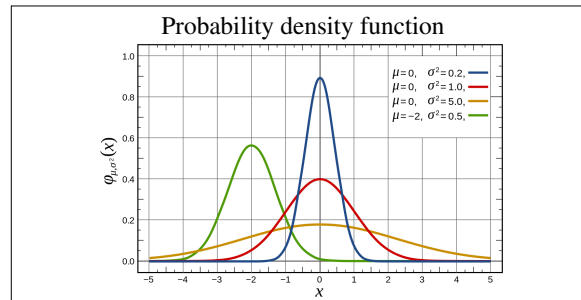
[1] Ternary plots (<http://wvaughan.org/ternaryplots.html>)

[2] <http://wvaughan.org/ternaryplots.html>

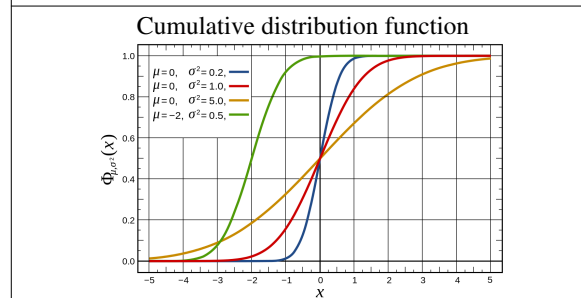
Distributions

Normal distribution

Normal



The red curve is the *standard normal distribution*



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
pdf	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x - \mu}{\sqrt{2\sigma^2}} \right) \right]$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
Skewness	0
Ex. kurtosis	0
Entropy	$\frac{1}{2} \ln(2\pi e \sigma^2)$
MGF	$\exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}$
CF	$\exp\left\{i\mu t - \frac{1}{2}\sigma^2 t^2\right\}$

Fisher information	$\begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$
---------------------------	---

In probability theory, the **normal** (or **Gaussian**) **distribution** is a very commonly occurring continuous probability distribution—a function that tells the probability that an observation in some context will fall between any two real numbers. For example, the distribution of grades on a test administered to many people is normally distributed. Normal distributions are extremely important in statistics and are often used in the natural and social sciences for real-valued random variables whose distributions are not known.^[1]

The normal distribution is immensely useful because of the central limit theorem, which states that, under mild conditions, the mean of many random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution: physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have a distribution very close to the normal. Moreover, many results and methods (such as propagation of uncertainty and least squares parameter fitting) can be derived analytically in explicit form when the relevant variables are normally distributed.

The Gaussian distribution is sometimes informally called the **bell curve**. However, many other distributions are bell-shaped (such as Cauchy's, Student's, and logistic). The terms **Gaussian function** and **Gaussian bell curve** are also ambiguous because they sometimes refer to multiples of the normal distribution that cannot be directly interpreted in terms of probabilities.

A normal distribution is

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The parameter μ in this definition is the *mean* or *expectation* of the distribution (and also its median and mode). The parameter σ is its standard deviation; its variance is therefore σ^2 . A random variable with a Gaussian distribution is said to be **normally distributed** and is called a **normal deviate**.

If $\mu = 0$ and $\sigma = 1$, the distribution is called the **standard normal distribution** or the **unit normal distribution**, and a random variable with that distribution is a **standard normal deviate**.

The normal distribution is the only absolutely continuous distribution all of whose cumulants beyond the first two (i.e., other than the mean and variance) are zero. It is also the continuous distribution with the maximum entropy for a given mean and variance.

The normal distribution is a subclass of the elliptical distributions. The normal distribution is symmetric about its mean, and is non-zero over the entire real line. As such it may not be a suitable model for variables that are inherently positive or strongly skewed, such as the weight of a person or the price of a share. Such variables may be better described by other distributions, such as the log-normal distribution or the Pareto distribution.

The value of the normal distribution is practically zero when the value x lies more than a few standard deviations away from the mean. Therefore, it may not be an appropriate model when one expects a significant fraction of outliers—values that lie many standard deviations away from the mean—and Least-squares and other statistical inference methods that are optimal for normally distributed variables often become highly unreliable when applied to such data. In those cases, assume a more heavy-tailed distribution and the appropriate robust statistical inference methods.

Definition

Standard normal distribution

The simplest case of a normal distribution is known as the *standard normal distribution*, described by this probability density function:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

The factor $1/\sqrt{2\pi}$ in this expression ensures that the total area under the curve $\phi(x)$ is equal to one^[proof]. The $1/2$ in the exponent ensures that the distribution has unit variance (and therefore also unit standard deviation). This function is symmetric around $x=0$, where it attains its maximum value $1/\sqrt{2\pi}$; and has inflection points at $+1$ and -1 .

General normal distribution

Any normal distribution is a version of the standard normal distribution whose domain has been stretched by a factor σ (the standard deviation) and then translated by μ (the mean value)

$$f(x, \mu, \sigma) = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right).$$

The probability density must be scaled by $1/\sigma$ so that the integral is still 1.

If Z is a standard normal deviate, then $X = Z\sigma + \mu$ will have a normal distribution with expected value μ and standard deviation σ . Conversely, if X is a general normal deviate, then $Z = (X - \mu)/\sigma$ will have a standard normal distribution.

Every normal distribution is the exponential of a quadratic function:

$$f(x) = e^{ax^2 + bx + c}$$

where a is negative and c is $-\ln(-4a\pi)/2$. In this form, the mean value μ is $-b/a$, and the variance σ^2 is $-1/(2a)$.

For the standard normal distribution, a is $-1/2$, b is zero, and c is $-\ln(2\pi)/2$.

Notation

The standard Gaussian distribution (with zero mean and unit variance) is often denoted with the Greek letter ϕ (phi). The alternative form of the Greek phi letter, φ , is also used quite often.

The normal distribution is also often denoted by $N(\mu, \sigma^2)$. Thus when a random variable X is distributed normally with mean μ and variance σ^2 , we write

$$X \sim N(\mu, \sigma^2).$$

Alternative parametrizations

Some authors advocate using the precision τ as the parameter defining the width of the distribution, instead of the deviation σ or the variance σ^2 . The precision is normally defined as the reciprocal of the variance, $1/\sigma^2$. The formula for the distribution then becomes

$$f(x) = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(x-\mu)^2}{2}}.$$

This choice is claimed to have advantages in numerical computations when σ is very close to zero and simplify formulas in some contexts, such as in the Bayesian inference of variables with multivariate normal distribution.

Occasionally, the precision τ is $1/\sigma$, the reciprocal of the standard deviation; so that

$$f(x) = \frac{\tau}{\sqrt{2\pi}} e^{-\frac{\tau^2(x-\mu)^2}{2}}.$$

Alternative definitions

Authors may differ also on which normal distribution should be called the "standard" one. Gauss himself defined the standard normal as having variance $\sigma^2 = 1/2$, that is

$$f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$

Stephen Stigler goes even further, defining the standard normal with variance $\sigma^2 = 1/2\pi$:

$$f(x) = e^{-\pi x^2}$$

According to Stigler, this formulation is advantageous because of a much simpler and easier-to-remember formula, the fact that the pdf has unit height at zero, and simple approximate formulas for the quantiles of the distribution.

Properties

Symmetries and derivatives

The normal distribution $f(x)$, with any mean μ and any positive deviation σ , has the following properties:

- It is symmetric around the point $x = \mu$, which is at the same time the mode, the median and the mean of the distribution.
- It is unimodal: its first derivative is positive for $x < \mu$, negative for $x > \mu$, and zero only at $x = \mu$.
- It has two inflection points (where the second derivative of f is zero and changes sign), located one standard deviation away from the mean, namely at $x = \mu - \sigma$ and $x = \mu + \sigma$.
- It is log-concave.
- It is infinitely differentiable, indeed supersmooth of order 2.

Furthermore, the standard normal distribution ϕ (with $\mu = 0$ and $\sigma = 1$) also has the following properties:

- Its first derivative $\phi'(x)$ is $-x\phi(x)$.
- Its second derivative $\phi''(x)$ is $(x^2 - 1)\phi(x)$
- More generally, its n -th derivative $\phi^{(n)}(x)$ is $(-1)^n H_n(x)\phi(x)$, where H_n is the Hermite polynomial of order n .

Moments

The plain and absolute moments of a variable X are the expected values of X^p and $|X|^p$, respectively. If the expected value μ of X is zero, these parameters are called *central moments*. Usually we are interested only in moments with integer order p .

If X has a normal distribution, these moments exist and are finite for any p whose real part is greater than -1 . For any non-negative integer p , the plain central moments are

$$\mathbb{E}[X^p] = \begin{cases} 0 & \text{if } p \text{ is odd,} \\ \sigma^p (p-1)!! & \text{if } p \text{ is even.} \end{cases}$$

Here $n!!$ denotes the double factorial, that is the product of every odd number from n to 1.

The central absolute moments coincide with plain moments for all even orders, but are nonzero for odd orders. For any non-negative integer p ,

$$\mathbb{E}[|X|^p] = \sigma^p (p-1)!! \cdot \begin{cases} \sqrt{\frac{2}{\pi}} & \text{if } p \text{ is odd} \\ 1 & \text{if } p \text{ is even} \end{cases} = \sigma^p \cdot \frac{2^{\frac{p}{2}} \Gamma(\frac{p+1}{2})}{\sqrt{\pi}}$$

The last formula is valid also for any non-integer $p > -1$.

When the mean μ is not zero, the plain and absolute moments can be expressed in terms of confluent hypergeometric functions ${}_1F_1$ and U .^[citation needed]

$$E [X^p] = \sigma^p \cdot (-i\sqrt{2} \operatorname{sgn} \mu)^p U \left(-\frac{1}{2}p, \frac{1}{2}, -\frac{1}{2}(\mu/\sigma)^2 \right),$$

$$E [|X|^p] = \sigma^p \cdot 2^{\frac{p}{2}} \frac{\Gamma \left(\frac{1+p}{2} \right)}{\sqrt{\pi}} {}_1F_1 \left(-\frac{1}{2}p, \frac{1}{2}, -\frac{1}{2}(\mu/\sigma)^2 \right).$$

These expressions remain valid even if p is not integer. See also generalized Hermite polynomials.

Order	Non-central moment	Central moment
1	μ	0
2	$\mu^2 + \sigma^2$	σ^2
3	$\mu^3 + 3\mu\sigma^2$	0
4	$\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$	$3\sigma^4$
5	$\mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4$	0
6	$\mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6$	$15\sigma^6$
7	$\mu^7 + 21\mu^5\sigma^2 + 105\mu^3\sigma^4 + 105\mu\sigma^6$	0
8	$\mu^8 + 28\mu^6\sigma^2 + 210\mu^4\sigma^4 + 420\mu^2\sigma^6 + 105\sigma^8$	$105\sigma^8$

Fourier transform and characteristic function

The Fourier transform of a normal distribution f with mean μ and deviation σ is

$$\hat{\phi}(t) = \int_{-\infty}^{\infty} f(x)e^{itx} dx = e^{i\mu t} e^{-\frac{1}{2}(\sigma t)^2}$$

where i is the imaginary unit. If the mean μ is zero, the first factor is 1, and the Fourier transform is also a normal distribution on the frequency domain, with mean 0 and standard deviation $1/\sigma$. In particular, the standard normal distribution ϕ (with $\mu=0$ and $\sigma=1$) is an eigenfunction of the Fourier transform.

In probability theory, the Fourier transform of the probability distribution of a real-valued random variable X is called the characteristic function of that variable, and can be defined as the expected value of e^{itX} , as a function of the real variable t (the frequency parameter of the Fourier transform). This definition can be analytically extended to a complex-value parameter t .

Moment and cumulant generating functions

The moment generating function of a real random variable X is the expected value of e^{tX} , as a function of the real parameter t . For a normal distribution with mean μ and deviation σ , the moment generating function exists and is equal to

$$M(t) = \hat{\phi}(-it) = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2}$$

The cumulant generating function is the logarithm of the moment generating function, namely

$$g(t) = \ln M(t) = \mu t + \frac{1}{2}\sigma^2 t^2$$

Since this is a quadratic polynomial in t , only the first two cumulants are nonzero, namely the mean μ and the variance σ^2 .

Cumulative distribution

The cumulative distribution function (CDF) of the standard normal distribution, usually denoted with the capital Greek letter Φ (phi), is the integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

Therefore here are some trivial results from area under bell curve -

$$\Phi(-\infty) = 0 = 0\%$$

$$\Phi(0) = 0.5 = 50\%$$

$$\Phi(\infty) = 1 = 100\%$$

$$\Phi(x) = 1 - \Phi(-x) \text{ and therefore } \Phi(x) + \Phi(-x) = 100\%$$

In statistics one often uses the related error function, or $\text{erf}(x)$, defined as the probability of a random variable with normal distribution of mean 0 and variance 1/2 falling in the range $[-x, x]$; that is

$$\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt$$

These integrals cannot be expressed in terms of elementary functions, and are often said to be special functions *. They are closely related, namely

$$\Phi(x) = \frac{1}{2} \left[1 + \text{erf} \left(\frac{x}{\sqrt{2}} \right) \right]$$

For a generic normal distribution f with mean μ and deviation σ , the cumulative distribution function is

$$F(x) = \Phi \left(\frac{x - \mu}{\sigma} \right) = \frac{1}{2} \left[1 + \text{erf} \left(\frac{x - \mu}{\sigma\sqrt{2}} \right) \right]$$

The complement of the standard normal CDF, $Q(x) = 1 - \Phi(x)$, is often called the Q-function, especially in engineering texts. It gives the probability that the value of a standard normal random variable X will exceed x . Other definitions of the Q-function, all of which are simple transformations of Φ , are also used occasionally.

The graph of the standard normal CDF Φ has 2-fold rotational symmetry around the point (0,1/2); that is, $\Phi(-x) = 1 - \Phi(x)$. Its antiderivative (indefinite integral) $\int \Phi(x) dx$ is $\int \Phi(x) dx = x\Phi(x) + \phi(x)$.

- The cumulative distribution function (CDF) of the standard normal distribution can be expanded by Integration by parts into a series:

$$\Phi(x) = 0.5 + \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \left[x + \frac{x^3}{3} + \frac{x^5}{3 \cdot 5} + \dots + \frac{x^{2n+1}}{3 \cdot 5 \cdot 7 \cdot \dots \cdot (2n+1)} \right]$$

Example of Pascal function to calculate CDF (sum of first 100 elements)

```
function CDF(x:extended):extended;
var value,sum:extended;
    i:integer;
begin
    sum:=x;
    value:=x;
    for i:=1 to 100 do
        begin
            value:=(value*x*x/(2*i+1));
            sum:=sum+value;
        end;
end;
```

```
result:=0.5+(sum/sqrt(2*pi))*exp(-(x*x)/2);
end;
```

Standard deviation and tolerance intervals

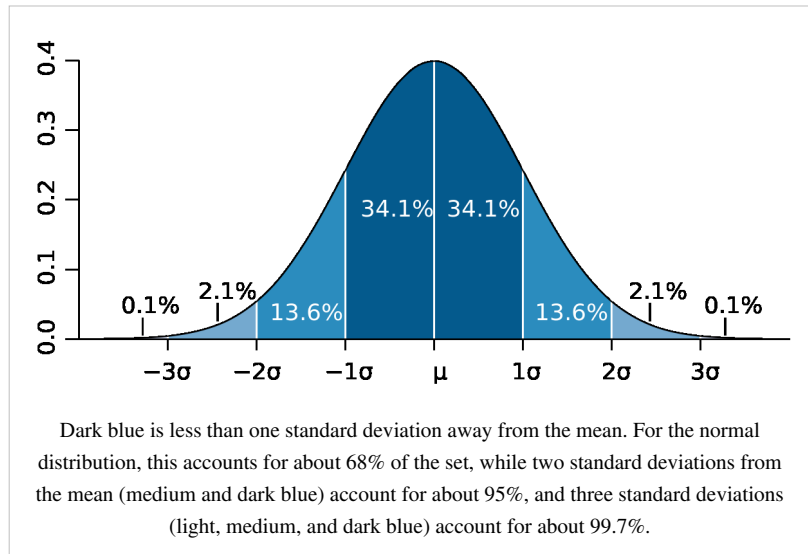
About 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations. This fact is known as the 68-95-99.7 (empirical) rule, or the 3-sigma rule.

More precisely, the probability that a normal deviate lies in the range $\mu - n\sigma$ and $\mu + n\sigma$ is given by

$$F(\mu + n\sigma) - F(\mu - n\sigma) = \Phi(n) - \Phi(-n) = \operatorname{erf}\left(\frac{n}{\sqrt{2}}\right),$$

To 12 decimal places, the values for $n = 1, 2, \dots, 6$ are:^[2]

n	$F(\mu+n\sigma) - F(\mu-n\sigma)$	i.e. 1 minus ...	or 1 in ...	OEIS
1	0.682689492137	0.317310507863	3.15148718753	A178647
2	0.954499736104	0.045500263896	21.9778945080	A110894
3	0.997300203937	0.002699796063	370.398347345	
4	0.999936657516	0.000063342484	15787.1927673	
5	0.999999426697	0.000000573303	1744277.89362	
6	0.99999998027	0.00000001973	506797345.897	



Quantile function

The quantile function of a distribution is the inverse of the cumulative distribution function. The quantile function of the standard normal distribution is called the probit function, and can be expressed in terms of the inverse error function:

$$\Phi^{-1}(p) = \sqrt{2} \operatorname{erf}^{-1}(2p - 1), \quad p \in (0, 1).$$

For a normal random variable with mean μ and variance σ^2 , the quantile function is

$$F^{-1}(p) = \mu + \sigma\Phi^{-1}(p) = \mu + \sigma\sqrt{2} \operatorname{erf}^{-1}(2p - 1), \quad p \in (0, 1).$$

The quantile $\Phi^{-1}(p)$ of the standard normal distribution is commonly denoted as z_p . These values are used in hypothesis testing, construction of confidence intervals and Q-Q plots. A normal random variable X will exceed $\mu + \sigma z_p$ with probability $1-p$; and will lie outside the interval $\mu \pm \sigma z_p$ with probability $2(1-p)$. In particular, the quantile $z_{0.975}$ is 1.96; therefore a normal random variable will lie outside the interval $\mu \pm 1.96\sigma$ in only 5% of cases. The following table gives the multiple n of σ such that X will lie in the range $\mu \pm n\sigma$ with a specified probability p . These values are useful to determine tolerance interval for sample averages and other statistical estimators with normal (or asymptotically normal) distributions:^[3]

$F(\mu+n\sigma) - F(\mu-n\sigma)$	n	$F(\mu+n\sigma) - F(\mu-n\sigma)$	n
0.80	1.281551565545	0.999	3.290526731492
0.90	1.644853626951	0.9999	3.890591886413
0.95	1.959963984540	0.99999	4.417173413469
0.98	2.326347874041	0.999999	4.891638475699
0.99	2.575829303549	0.9999999	5.326723886384
0.995	2.807033768344	0.99999999	5.730728868236
0.998	3.090232306168	0.999999999	6.109410204869

Zero-variance limit

In the limit when σ tends to zero, the probability density $f(x)$ eventually tends to zero at any $x \neq \mu$, but grows without limit if $x = \mu$, while its integral remains equal to 1. Therefore, the normal distribution cannot be defined as an ordinary function when $\sigma = 0$.

However, one can define the normal distribution with zero variance as a generalized function; specifically, as Dirac's "delta function" δ translated by the mean μ , that is $f(x) = \delta(x-\mu)$. Its CDF is then the Heaviside step function translated by the mean μ , namely

$$F(x) = \begin{cases} 0 & \text{if } x < \mu \\ 1 & \text{if } x \geq \mu \end{cases}$$

The central limit theorem

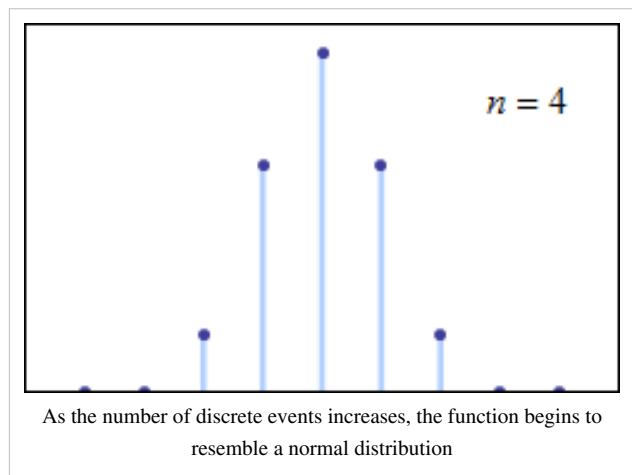
The central limit theorem states that under certain (fairly common) conditions, the sum of many random variables will have an approximately normal distribution. More specifically, where X_1, \dots, X_n are independent and identically distributed random variables with the same arbitrary distribution, zero mean, and variance σ^2 ; and Z is their mean scaled by \sqrt{n}

$$Z = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n X_i \right)$$

Then, as n increases, the probability distribution of Z will tend to the normal distribution with zero mean and variance σ^2 .

The theorem can be extended to variables X_i that are not independent and/or not identically distributed if certain constraints are placed on the degree of dependence and the moments of the distributions.

Many test statistics, scores, and estimators encountered in practice contain sums of certain random variables in

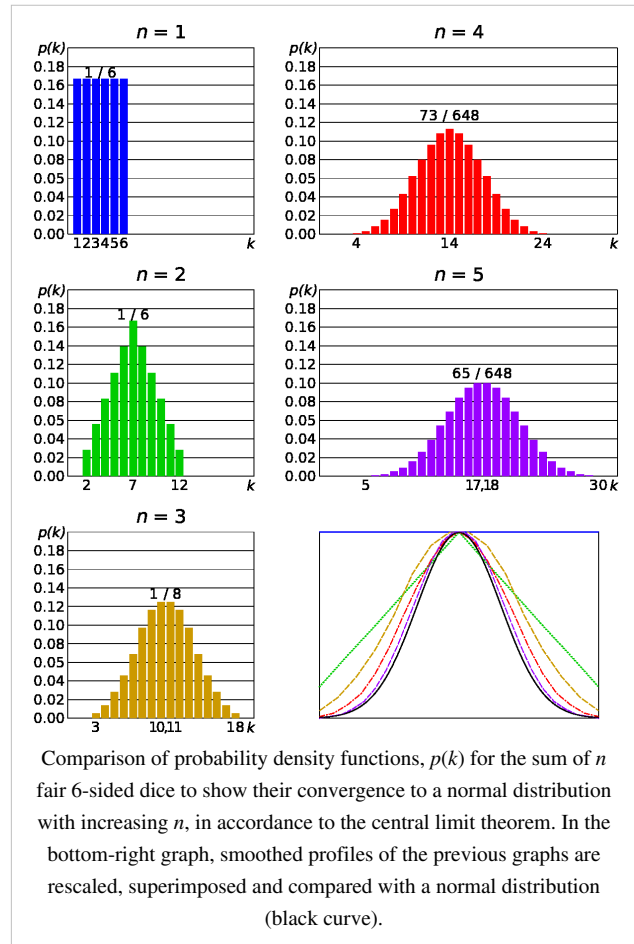


them, and even more estimators can be represented as sums of random variables through the use of influence functions. The central limit theorem implies that those statistical parameters will have asymptotically normal distributions.

The central limit theorem also implies that certain distributions can be approximated by the normal distribution, for example:

- The binomial distribution $B(n, p)$ is approximately normal with mean np and variance $np(1-p)$ for large n and for p not too close to zero or one.
- The Poisson distribution with parameter λ is approximately normal with mean λ and variance λ , for large values of λ .^[4]
- The chi-squared distribution $\chi^2(k)$ is approximately normal with mean k and variance $2k$, for large k .
- The Student's t-distribution $t(v)$ is approximately normal with mean 0 and variance 1 when v is large.

Whether these approximations are sufficiently accurate depends on the purpose for which they are needed, and the rate of convergence to the normal distribution. It is typically the case that such approximations are less accurate in the tails of the distribution.



A general upper bound for the approximation error in the central limit theorem is given by the Berry–Esseen theorem, improvements of the approximation are given by the Edgeworth expansions.

Operations on normal deviates

The family of normal distributions is closed under linear transformations: if X is normally distributed with mean μ and deviation σ , then the variable $Y = aX + b$, for any real numbers a and b , is also normally distributed, with mean $a\mu + b$ and deviation $a\sigma$.

Also if X_1 and X_2 are two independent normal random variables, with means μ_1, μ_2 and standard deviations σ_1, σ_2 , then their sum $X_1 + X_2$ will also be normally distributed,^[proof] with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

In particular, if X and Y are independent normal deviates with zero mean and variance σ^2 , then $X + Y$ and $X - Y$ are also independent and normally distributed, with zero mean and variance $2\sigma^2$. This is a special case of the polarization identity.

Also, if X_1, X_2 are two independent normal deviates with mean μ and deviation σ , and a, b are arbitrary real numbers, then the variable

$$X_3 = \frac{aX_1 + bX_2 - (a + b)\mu}{\sqrt{a^2 + b^2}} + \mu$$

is also normally distributed with mean μ and deviation σ . It follows that the normal distribution is stable (with exponent $\alpha = 2$).

More generally, any linear combination of independent normal deviates is a normal deviate.

Infinite divisibility and Cramér's theorem

For any positive integer n , any normal distribution with mean μ and variance σ^2 is the distribution of the sum of n independent normal deviates, each with mean μ/n and variance σ^2/n . This property is called infinite divisibility.

Conversely, if X_1 and X_2 are independent random variables and their sum $X_1 + X_2$ has a normal distribution, then both X_1 and X_2 must be normal deviates.

This result is known as **Cramér's decomposition theorem**, and is equivalent to saying that the convolution of two distributions is normal if and only if both are normal. Cramér's theorem implies that a linear combination of independent non-Gaussian variables will never have an exactly normal distribution, although it may approach it arbitrarily close.

Bernstein's theorem

Bernstein's theorem states that if X and Y are independent and $X + Y$ and $X - Y$ are also independent, then both X and Y must necessarily have normal distributions.^[5]

More generally, if X_1, \dots, X_n are independent random variables, then two distinct linear combinations $\sum a_k X_k$ and $\sum b_k X_k$ will be independent if and only if all X_k 's are normal and $\sum a_k b_k \sigma_k^2 = 0$, where σ_k^2

σ_k^2 denotes the variance of X_k .

Other properties

1. If the characteristic function φ_X of some random variable X is of the form $\varphi_X(t) = e^{Q(t)}$, where $Q(t)$ is a polynomial, then the **Marcinkiewicz theorem** (named after Józef Marcinkiewicz) asserts that Q can be at most a quadratic polynomial, and therefore X a normal random variable. The consequence of this result is that the normal distribution is the only distribution with a finite number (two) of non-zero cumulants.
2. If X and Y are jointly normal and uncorrelated, then they are independent. The requirement that X and Y should be jointly normal is essential, without it the property does not hold.^{[6][7][proof]} For non-normal random variables uncorrelatedness does not imply independence.
3. The Kullback–Leibler divergence of one normal distributions $X_1 \sim N(\mu_1, \sigma_1^2)$ from another $X_2 \sim N(\mu_2, \sigma_2^2)$ is given by:^[8]

$$D_{\text{KL}}(X_1 \parallel X_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} + \frac{1}{2} \left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \ln \frac{\sigma_1^2}{\sigma_2^2} \right).$$

The Hellinger distance between the same distributions is equal to

$$H^2(X_1, X_2) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} e^{-\frac{1}{4} \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}}.$$

4. The Fisher information matrix for a normal distribution is diagonal and takes the form

$$\mathcal{I} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

5. Normal distributions belongs to an exponential family with natural parameters $\theta_1 = \frac{\mu}{\sigma^2}$ and $\theta_2 = \frac{-1}{2\sigma^2}$, and natural statistics x and x^2 . The dual, expectation parameters for normal distribution are $\eta_1 = \mu$ and $\eta_2 = \mu^2 + \sigma^2$.
6. The conjugate prior of the mean of a normal distribution is another normal distribution. Specifically, if x_1, \dots, x_n are iid $N(\mu, \sigma^2)$ and the prior is $\mu \sim N(\mu_0, \sigma_0^2)$, then the posterior distribution for the estimator of μ will be

$$\mu | x_1, \dots, x_n \sim \mathcal{N} \left(\frac{\frac{\sigma^2}{n} \mu_0 + \sigma_0^2 \bar{x}}{\frac{\sigma^2}{n} + \sigma_0^2}, \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1} \right)$$

7. Of all probability distributions over the reals with mean μ and variance σ^2 , the normal distribution $N(\mu, \sigma^2)$ is the one with the maximum entropy.
8. The family of normal distributions forms a manifold with constant curvature -1 . The same family is flat with respect to the (± 1) -connections $\nabla^{(e)}$ and $\nabla^{(m)}$.

Related distributions

Operations on a single random variable

If X is distributed normally with mean μ and variance σ^2 , then

- The exponential of X is distributed log-normally: $e^X \sim \ln(N(\mu, \sigma^2))$.
- The absolute value of X has folded normal distribution: $|X| \sim N_f(\mu, \sigma^2)$. If $\mu = 0$ this is known as the half-normal distribution.
- The square of X/σ has the noncentral chi-squared distribution with one degree of freedom: $X^2/\sigma^2 \sim \chi^2_1(\mu^2/\sigma^2)$. If $\mu = 0$, the distribution is called simply chi-squared.
- The distribution of the variable X restricted to an interval $[a, b]$ is called the truncated normal distribution.
- $(X - \mu)^{-2}$ has a Lévy distribution with location 0 and scale σ^{-2} .

Combination of two independent random variables

If X_1 and X_2 are two independent standard normal random variables with mean 0 and variance 1, then

- Their sum and difference is distributed normally with mean zero and variance two: $X_1 \pm X_2 \sim N(0, 2)$.
- Their product $Z = X_1 \cdot X_2$ follows the "product-normal" distribution^[9] with density function $f_Z(z) = \pi^{-1} K_0(|z|)$, where K_0 is the modified Bessel function of the second kind. This distribution is symmetric around zero, unbounded at $z = 0$, and has the characteristic function $\varphi_Z(t) = (1 + t^2)^{-1/2}$.
- Their ratio follows the standard Cauchy distribution: $X_1 \div X_2 \sim \text{Cauchy}(0, 1)$.
- Their Euclidean norm $\sqrt{X_1^2 + X_2^2}$ has the Rayleigh distribution.

Combination of two or more independent random variables

- If X_1, X_2, \dots, X_n are independent standard normal random variables, then the sum of their squares has the chi-squared distribution with n degrees of freedom

$$X_1^2 + \dots + X_n^2 \sim \chi_n^2.$$

- If X_1, X_2, \dots, X_n are independent normally distributed random variables with means μ and variances σ^2 , then their sample mean is independent from the sample standard deviation, which can be demonstrated using Basu's theorem or Cochran's theorem. The ratio of these two quantities will have the Student's t-distribution with $n - 1$ degrees of freedom:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\frac{1}{n}(X_1 + \dots + X_n) - \mu}{\sqrt{\frac{1}{n(n-1)} [(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2]}} \sim t_{n-1}.$$

- If $X_1, \dots, X_n, Y_1, \dots, Y_m$ are independent standard normal random variables, then the ratio of their normalized sums of squares will have the F-distribution with (n, m) degrees of freedom:

$$F = \frac{(X_1^2 + X_2^2 + \dots + X_n^2)/n}{(Y_1^2 + Y_2^2 + \dots + Y_m^2)/m} \sim F_{n,m}.$$

Operations on the density function

The split normal distribution is most directly defined in terms of joining scaled sections of the density functions of different normal distributions and rescaling the density to integrate to one. The truncated normal distribution results from rescaling a section of a single density function.

Extensions

The notion of normal distribution, being one of the most important distributions in probability theory, has been extended far beyond the standard framework of the univariate (that is one-dimensional) case (Case 1). All these extensions are also called *normal* or *Gaussian* laws, so a certain ambiguity in names exists.

- The multivariate normal distribution describes the Gaussian law in the k -dimensional Euclidean space. A vector $X \in \mathbf{R}^k$ is multivariate-normally distributed if any linear combination of its components $\sum_{j=1}^k a_j X_j$ has a (univariate) normal distribution. The variance of X is a $k \times k$ symmetric positive-definite matrix V . The multivariate normal distribution is a special case of the elliptical distributions. As such, its iso-density loci in the $k = 2$ case are ellipses and in the case of arbitrary k are ellipsoids.
- Rectified Gaussian distribution a rectified version of normal distribution with all the negative elements reset to 0
- Complex normal distribution deals with the complex normal vectors. A complex vector $X \in \mathbf{C}^k$ is said to be normal if both its real and imaginary components jointly possess a $2k$ -dimensional multivariate normal distribution. The variance-covariance structure of X is described by two matrices: the *variance* matrix Γ , and the *relation* matrix C .
- Matrix normal distribution describes the case of normally distributed matrices.
- Gaussian processes are the normally distributed stochastic processes. These can be viewed as elements of some infinite-dimensional Hilbert space H , and thus are the analogues of multivariate normal vectors for the case $k = \infty$. A random element $h \in H$ is said to be normal if for any constant $a \in H$ the scalar product (a, h) has a (univariate) normal distribution. The variance structure of such Gaussian random element can be described in terms of the linear *covariance operator* $K: H \rightarrow H$. Several Gaussian processes became popular enough to have their own names:
 - Brownian motion,
 - Brownian bridge,
 - Ornstein–Uhlenbeck process.
- Gaussian q -distribution is an abstract mathematical construction that represents a " q -analogue" of the normal distribution.
- the q -Gaussian is an analogue of the Gaussian distribution, in the sense that it maximises the Tsallis entropy, and is one type of Tsallis distribution. Note that this distribution is different from the Gaussian q -distribution above.

One of the main practical uses of the Gaussian law is to model the empirical distributions of many different random variables encountered in practice. In such case a possible extension would be a richer family of distributions, having more than two parameters and therefore being able to fit the empirical distribution more accurately. The examples of such extensions are:

- Pearson distribution— a four-parametric family of probability distributions that extend the normal law to include different skewness and kurtosis values.

Normality tests

Normality tests assess the likelihood that the given data set $\{x_1, \dots, x_n\}$ comes from a normal distribution. Typically the null hypothesis H_0 is that the observations are distributed normally with unspecified mean μ and variance σ^2 , versus the alternative H_a that the distribution is arbitrary. Many tests (over 40) have been devised for this problem, the more prominent of them are outlined below:

- **"Visual" tests** are more intuitively appealing but subjective at the same time, as they rely on informal human judgement to accept or reject the null hypothesis.
 - Q-Q plot— is a plot of the sorted values from the data set against the expected values of the corresponding quantiles from the standard normal distribution. That is, it's a plot of point of the form $(\Phi^{-1}(p_k), x_{(k)})$, where plotting points p_k are equal to $p_k = (k - \alpha)/(n + 1 - 2\alpha)$ and α is an adjustment constant, which can be anything between 0 and 1. If the null hypothesis is true, the plotted points should approximately lie on a straight line.
 - P-P plot— similar to the Q-Q plot, but used much less frequently. This method consists of plotting the points $(\Phi(z_{(k)}), p_k)$, where $z_{(k)} = (x_{(k)} - \hat{\mu})/\hat{\sigma}$. For normally distributed data this plot should lie on a 45° line between $(0, 0)$ and $(1, 1)$.
 - Shapiro-Wilk test employs the fact that the line in the Q-Q plot has the slope of σ . The test compares the least squares estimate of that slope with the value of the sample variance, and rejects the null hypothesis if these two quantities differ significantly.
 - Normal probability plot (rankit plot)
- **Moment tests:**
 - D'Agostino's K-squared test
 - Jarque–Bera test
- **Empirical distribution function tests:**
 - Lilliefors test (an adaptation of the Kolmogorov–Smirnov test)
 - Anderson–Darling test

Estimation of parameters

It is often the case that we don't know the parameters of the normal distribution, but instead want to estimate them. That is, having a sample (x_1, \dots, x_n) from a normal $N(\mu, \sigma^2)$ population we would like to learn the approximate values of parameters μ and σ^2 . The standard approach to this problem is the maximum likelihood method, which requires maximization of the *log-likelihood function*:

$$\ln \mathcal{L}(\mu, \sigma^2) = \sum_{i=1}^n \ln f(x_i; \mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Taking derivatives with respect to μ and σ^2 and solving the resulting system of first order conditions yields the *maximum likelihood estimates*:

$$\hat{\mu} = \bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Estimator $\hat{\mu}$ is called the *sample mean*, since it is the arithmetic mean of all observations. The statistic \bar{x} is complete and sufficient for μ , and therefore by the Lehmann–Scheffé theorem, $\hat{\mu}$ is the uniformly minimum variance unbiased (UMVU) estimator. In finite samples it is distributed normally:

$$\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/n).$$

The variance of this estimator is equal to the $\mu\mu$ -element of the inverse Fisher information matrix \mathcal{I}^{-1} . This implies that the estimator is finite-sample efficient. Of practical importance is the fact that the standard error of $\hat{\mu}$ is proportional to $1/\sqrt{n}$, that is, if one wishes to decrease the standard error by a factor of 10, one must increase the number of points in the sample by a factor of 100. This fact is widely used in determining sample sizes for opinion

polls and the number of trials in Monte Carlo simulations.

From the standpoint of the asymptotic theory, $\hat{\mu}$ is consistent, that is, it converges in probability to μ as $n \rightarrow \infty$. The estimator is also asymptotically normal, which is a simple corollary of the fact that it is normal in finite samples:

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

The estimator $\hat{\sigma}^2$ is called the *sample variance*, since it is the variance of the sample (x_1, \dots, x_n) . In practice, another estimator is often used instead of the $\hat{\sigma}^2$. This other estimator is denoted s^2 , and is also called the *sample variance*, which represents a certain ambiguity in terminology; its square root s is called the *sample standard deviation*. The estimator s^2 differs from $\hat{\sigma}^2$ by having $(n - 1)$ instead of n in the denominator (the so-called Bessel's correction):

$$s^2 = \frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

The difference between s^2 and $\hat{\sigma}^2$ becomes negligibly small for large n 's. In finite samples however, the motivation behind the use of s^2 is that it is an unbiased estimator of the underlying parameter σ^2 , whereas $\hat{\sigma}^2$ is biased. Also, by the Lehmann–Scheffé theorem the estimator s^2 is uniformly minimum variance unbiased (UMVU), which makes it the "best" estimator among all unbiased ones. However it can be shown that the biased estimator $\hat{\sigma}^2$ is "better" than the s^2 in terms of the mean squared error (MSE) criterion. In finite samples both s^2 and $\hat{\sigma}^2$ have scaled chi-squared distribution with $(n - 1)$ degrees of freedom:

$$s^2 \sim \frac{\sigma^2}{n-1} \cdot \chi_{n-1}^2, \quad \hat{\sigma}^2 \sim \frac{\sigma^2}{n} \cdot \chi_{n-1}^2.$$

The first of these expressions shows that the variance of s^2 is equal to $2\sigma^4/(n-1)$, which is slightly greater than the $\sigma\sigma$ -element of the inverse Fisher information matrix \mathcal{I}^{-1} . Thus, s^2 is not an efficient estimator for σ^2 , and moreover, since s^2 is UMVU, we can conclude that the finite-sample efficient estimator for σ^2 does not exist.

Applying the asymptotic theory, both estimators s^2 and $\hat{\sigma}^2$ are consistent, that is they converge in probability to σ^2 as the sample size $n \rightarrow \infty$. The two estimators are also both asymptotically normal:

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \simeq \sqrt{n}(s^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2\sigma^4).$$

In particular, both estimators are asymptotically efficient for σ^2 .

By Cochran's theorem, for normal distributions the sample mean $\hat{\mu}$ and the sample variance s^2 are independent, which means there can be no gain in considering their joint distribution. There is also a reverse theorem: if in a sample the sample mean and sample variance are independent, then the sample must have come from the normal distribution. The independence between $\hat{\mu}$ and s can be employed to construct the so-called *t-statistic*:

$$t = \frac{\hat{\mu} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - \mu}{\sqrt{\frac{1}{n(n-1)} \sum (x_i - \bar{x})^2}} \sim t_{n-1}$$

This quantity t has the Student's t-distribution with $(n - 1)$ degrees of freedom, and it is an ancillary statistic (independent of the value of the parameters). Inverting the distribution of this t -statistics will allow us to construct the confidence interval for μ ; similarly, inverting the χ^2 distribution of the statistic s^2 will give us the confidence interval for σ^2 :

$$\mu \in \left[\hat{\mu} + t_{n-1, \alpha/2} \frac{1}{\sqrt{n}} s, \hat{\mu} + t_{n-1, 1-\alpha/2} \frac{1}{\sqrt{n}} s \right] \approx \left[\hat{\mu} - |z_{\alpha/2}| \frac{1}{\sqrt{n}} s, \hat{\mu} + |z_{\alpha/2}| \frac{1}{\sqrt{n}} s \right],$$

$$\sigma^2 \in \left[\frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2} \right] \approx \left[s^2 - |z_{\alpha/2}| \frac{\sqrt{2}}{\sqrt{n}} s^2, s^2 + |z_{\alpha/2}| \frac{\sqrt{2}}{\sqrt{n}} s^2 \right],$$

where $t_{k,p}$ and χ^2

k,p are the p^{th} quantiles of the t - and χ^2 -distributions respectively. These confidence intervals are of the level $1 - \alpha$, meaning that the true values μ and σ^2 fall outside of these intervals with probability α . In practice people usually take $\alpha = 5\%$, resulting in the 95% confidence intervals. The approximate formulas in the display above were derived from

the asymptotic distributions of $\hat{\mu}$ and s^2 . The approximate formulas become valid for large values of n , and are more convenient for the manual calculation since the standard normal quantiles $z_{\alpha/2}$ do not depend on n . In particular, the most popular value of $\alpha = 5\%$, results in $|z_{0.025}| = 1.96$.

Bayesian analysis of the normal distribution

Bayesian analysis of normally distributed data is complicated by the many different possibilities that may be considered:

- Either the mean, or the variance, or neither, may be considered a fixed quantity.
- When the variance is unknown, analysis may be done directly in terms of the variance, or in terms of the precision, the reciprocal of the variance. The reason for expressing the formulas in terms of precision is that the analysis of most cases is simplified.
- Both univariate and multivariate cases need to be considered.
- Either conjugate or improper prior distributions may be placed on the unknown variables.
- An additional set of cases occurs in Bayesian linear regression, where in the basic model the data is assumed to be normally distributed, and normal priors are placed on the regression coefficients. The resulting analysis is similar to the basic cases of independent identically distributed data, but more complex.

The formulas for the non-linear-regression cases are summarized in the conjugate prior article.

The sum of two quadratics

Scalar form

The following auxiliary formula is useful for simplifying the posterior update equations, which otherwise become fairly tedious.

$$a(x - y)^2 + b(x - z)^2 = (a + b) \left(x - \frac{ay + bz}{a + b} \right)^2 + \frac{ab}{a + b} (y - z)^2$$

This equation rewrites the sum of two quadratics in x by expanding the squares, grouping the terms in x , and completing the square. Note the following about the complex constant factors attached to some of the terms:

1. The factor $\frac{ay + bz}{a + b}$ has the form of a weighted average of y and z .
2. $\frac{ab}{a + b} = \frac{1}{\frac{1}{a} + \frac{1}{b}} = (a^{-1} + b^{-1})^{-1}$. This shows that this factor can be thought of as resulting from a

situation where the reciprocals of quantities a and b add directly, so to combine a and b themselves, it's necessary to reciprocate, add, and reciprocate the result again to get back into the original units. This is exactly the sort of operation performed by the harmonic mean, so it is not surprising that $\frac{ab}{a + b}$ is one-half the harmonic mean of a and b .

Vector form

A similar formula can be written for the sum of two vector quadratics: If \mathbf{x} , \mathbf{y} , \mathbf{z} are vectors of length k , and \mathbf{A} and \mathbf{B} are symmetric, invertible matrices of size $k \times k$, then

$$(\mathbf{y}-\mathbf{x})'\mathbf{A}(\mathbf{y}-\mathbf{x})+(\mathbf{x}-\mathbf{z})'\mathbf{B}(\mathbf{x}-\mathbf{z})=(\mathbf{x}-\mathbf{c})'(\mathbf{A}+\mathbf{B})(\mathbf{x}-\mathbf{c})+(\mathbf{y}-\mathbf{z})'(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{y}-\mathbf{z})$$

where

$$\mathbf{c}=(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}\mathbf{y}+\mathbf{B}\mathbf{z})$$

Note that the form $\mathbf{x}'\mathbf{A}\mathbf{x}$ is called a quadratic form and is a scalar:

$$\mathbf{x}'\mathbf{A}\mathbf{x}=\sum_{i,j}a_{ij}x_ix_j$$

In other words, it sums up all possible combinations of products of pairs of elements from \mathbf{x} , with a separate coefficient for each. In addition, since $x_ix_j=x_jx_i$, only the sum $a_{ij}+a_{ji}$ matters for any off-diagonal elements of \mathbf{A} , and there is no loss of generality in assuming that \mathbf{A} is symmetric. Furthermore, if \mathbf{A} is symmetric, then the form $\mathbf{x}'\mathbf{A}\mathbf{y}=\mathbf{y}'\mathbf{A}\mathbf{x}$.

The sum of differences from the mean

Another useful formula is as follows:

$$\sum_{i=1}^n(x_i-\mu)^2=\sum_{i=1}^n(x_i-\bar{x})^2+n(\bar{x}-\mu)^2$$

where $\bar{x}=\frac{1}{n}\sum_{i=1}^nx_i$.

With known variance

For a set of i.i.d. normally distributed data points \mathbf{X} of size n where each individual point x follows $x\sim\mathcal{N}(\mu,\sigma^2)$ with known variance σ^2 , the conjugate prior distribution is also normally distributed.

This can be shown more easily by rewriting the variance as the precision, i.e. using $\tau=1/\sigma^2$. Then if $x\sim\mathcal{N}(\mu,\tau)$ and $\mu\sim\mathcal{N}(\mu_0,\tau_0)$, we proceed as follows.

First, the likelihood function is (using the formula above for the sum of differences from the mean):

$$\begin{aligned} p(\mathbf{X}|\mu,\tau) &= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x_i-\mu)^2\right) \\ &= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2}\tau\sum_{i=1}^n(x_i-\mu)^2\right) \\ &= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{2}\tau\left(\sum_{i=1}^n(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right)\right]. \end{aligned}$$

Then, we proceed as follows:

$$\begin{aligned}
p(\mu|\mathbf{X}) &\propto p(\mathbf{X}|\mu)p(\mu) \\
&= \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{2}\tau\left(\sum_{i=1}^n(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)\right] \sqrt{\frac{\tau_0}{2\pi}} \exp\left(-\frac{1}{2}\tau_0(\mu - \mu_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\tau\left(\sum_{i=1}^n(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right) + \tau_0(\mu - \mu_0)^2\right)\right) \\
&\propto \exp\left(-\frac{1}{2}(n\tau(\bar{x} - \mu)^2 + \tau_0(\mu - \mu_0)^2)\right) \\
&= \exp\left(-\frac{1}{2}(n\tau + \tau_0)\left(\mu - \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}\right)^2 + \frac{n\tau\tau_0}{n\tau + \tau_0}(\bar{x} - \mu_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2}(n\tau + \tau_0)\left(\mu - \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}\right)^2\right)
\end{aligned}$$

In the above derivation, we used the formula above for the sum of two quadratics and eliminated all constant factors not involving μ . The result is the kernel of a normal distribution, with mean $\frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}$ and precision $n\tau + \tau_0$,

i.e.

$$p(\mu|\mathbf{X}) \sim \mathcal{N}\left(\frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}, n\tau + \tau_0\right)$$

This can be written as a set of Bayesian update equations for the posterior parameters in terms of the prior parameters:

$$\begin{aligned}
\tau'_0 &= \tau_0 + n\tau \\
\mu'_0 &= \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0} \\
\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i
\end{aligned}$$

That is, to combine n data points with total precision of $n\tau$ (or equivalently, total variance of n/σ^2) and mean of values \bar{x} , derive a new total precision simply by adding the total precision of the data to the prior total precision, and form a new mean through a *precision-weighted average*, i.e. a weighted average of the data mean and the prior mean, each weighted by the associated total precision. This makes logical sense if the precision is thought of as indicating the certainty of the observations: In the distribution of the posterior mean, each of the input components is weighted by its certainty, and the certainty of this distribution is the sum of the individual certainties. (For the intuition of this, compare the expression "the whole is (or is not) greater than the sum of its parts". In addition, consider that the knowledge of the posterior comes from a combination of the knowledge of the prior and likelihood, so it makes sense that we are more certain of it than of either of its components.)

The above formula reveals why it is more convenient to do Bayesian analysis of conjugate priors for the normal distribution in terms of the precision. The posterior precision is simply the sum of the prior and likelihood precisions, and the posterior mean is computed through a precision-weighted average, as described above. The same formulas can be written in terms of variance by reciprocating all the precisions, yielding the more ugly formulas

$$\begin{aligned}\sigma_0^{2'} &= \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \\ \mu_0' &= \frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \\ \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i\end{aligned}$$

With known mean

For a set of i.i.d. normally distributed data points \mathbf{X} of size n where each individual point x follows $x \sim \mathcal{N}(\mu, \sigma^2)$ with known mean μ , the conjugate prior of the variance has an inverse gamma distribution or a scaled inverse chi-squared distribution. The two are equivalent except for having different parameterizations. Although the inverse gamma is more commonly used, we use the scaled inverse chi-squared for the sake of convenience. The prior for σ^2 is as follows:

$$p(\sigma^2 | \nu_0, \sigma_0^2) = \frac{(\sigma_0^2 \frac{\nu_0}{2})^{\frac{\nu_0}{2}} \exp\left[\frac{-\nu_0 \sigma_0^2}{2\sigma^2}\right]}{\Gamma\left(\frac{\nu_0}{2}\right) (\sigma^2)^{1+\frac{\nu_0}{2}}} \propto \frac{\exp\left[\frac{-\nu_0 \sigma_0^2}{2\sigma^2}\right]}{(\sigma^2)^{1+\frac{\nu_0}{2}}}$$

The likelihood function from above, written in terms of the variance, is:

$$\begin{aligned}p(\mathbf{X} | \mu, \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right] \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{S}{2\sigma^2}\right]\end{aligned}$$

where

$$S = \sum_{i=1}^n (x_i - \mu)^2.$$

Then:

$$\begin{aligned}p(\sigma^2 | \mathbf{X}) &\propto p(\mathbf{X} | \sigma^2) p(\sigma^2) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{S}{2\sigma^2}\right] \frac{(\sigma_0^2 \frac{\nu_0}{2})^{\frac{\nu_0}{2}} \exp\left[\frac{-\nu_0 \sigma_0^2}{2\sigma^2}\right]}{\Gamma\left(\frac{\nu_0}{2}\right) (\sigma^2)^{1+\frac{\nu_0}{2}}} \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \frac{1}{(\sigma^2)^{1+\frac{\nu_0}{2}}} \exp\left[-\frac{S}{2\sigma^2} + \frac{-\nu_0 \sigma_0^2}{2\sigma^2}\right] \\ &= \frac{1}{(\sigma^2)^{1+\frac{\nu_0+n}{2}}} \exp\left[-\frac{\nu_0 \sigma_0^2 + S}{2\sigma^2}\right]\end{aligned}$$

The above is also a scaled inverse chi-squared distribution where

$$\begin{aligned}\nu_0' &= \nu_0 + n \\ \nu_0' \sigma_0^{2'} &= \nu_0 \sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2\end{aligned}$$

or equivalently

$$\begin{aligned}\nu_0' &= \nu_0 + n \\ \sigma_0^{2'} &= \frac{\nu_0 \sigma_0^2 + \sum_{i=1}^n (x_i - \mu)^2}{\nu_0 + n}\end{aligned}$$

Reparameterizing in terms of an inverse gamma distribution, the result is:

$$\alpha' = \alpha + \frac{n}{2}$$

$$\beta' = \beta + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}$$

With unknown mean and unknown variance

For a set of i.i.d. normally distributed data points \mathbf{X} of size n where each individual point x follows $x \sim \mathcal{N}(\mu, \sigma^2)$ with unknown mean μ and unknown variance σ^2 , a combined (multivariate) conjugate prior is placed over the mean and variance, consisting of a normal-inverse-gamma distribution. Logically, this originates as follows:

1. From the analysis of the case with unknown mean but known variance, we see that the update equations involve sufficient statistics computed from the data consisting of the mean of the data points and the total variance of the data points, computed in turn from the known variance divided by the number of data points.
2. From the analysis of the case with unknown variance but known mean, we see that the update equations involve sufficient statistics over the data consisting of the number of data points and sum of squared deviations.
3. Keep in mind that the posterior update values serve as the prior distribution when further data is handled. Thus, we should logically think of our priors in terms of the sufficient statistics just described, with the same semantics kept in mind as much as possible.
4. To handle the case where both mean and variance are unknown, we could place independent priors over the mean and variance, with fixed estimates of the average mean, total variance, number of data points used to compute the variance prior, and sum of squared deviations. Note however that in reality, the total variance of the mean depends on the unknown variance, and the sum of squared deviations that goes into the variance prior (appears to) depend on the unknown mean. In practice, the latter dependence is relatively unimportant: Shifting the actual mean shifts the generated points by an equal amount, and on average the squared deviations will remain the same. This is not the case, however, with the total variance of the mean: As the unknown variance increases, the total variance of the mean will increase proportionately, and we would like to capture this dependence.
5. This suggests that we create a *conditional prior* of the mean on the unknown variance, with a hyperparameter specifying the mean of the pseudo-observations associated with the prior, and another parameter specifying the number of pseudo-observations. This number serves as a scaling parameter on the variance, making it possible to control the overall variance of the mean relative to the actual variance parameter. The prior for the variance also has two hyperparameters, one specifying the sum of squared deviations of the pseudo-observations associated with the prior, and another specifying once again the number of pseudo-observations. Note that each of the priors has a hyperparameter specifying the number of pseudo-observations, and in each case this controls the relative variance of that prior. These are given as two separate hyperparameters so that the variance (aka the confidence) of the two priors can be controlled separately.
6. This leads immediately to the normal-inverse-gamma distribution, which is the product of the two distributions just defined, with conjugate priors used (an inverse gamma distribution over the variance, and a normal distribution over the mean, *conditional* on the variance) and with the same four parameters just defined.

The priors are normally defined as follows:

$$p(\mu|\sigma^2; \mu_0, n_0) \sim \mathcal{N}(\mu_0, \sigma^2/n_0)$$

$$p(\sigma^2; \nu_0, \sigma_0^2) \sim I\chi^2(\nu_0, \sigma_0^2) = IG(\nu_0/2, \nu_0\sigma_0^2/2)$$

The update equations can be derived, and look as follows:

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ \mu'_0 &= \frac{n_0 \mu_0 + n \bar{x}}{n_0 + n} \\ n'_0 &= n_0 + n \\ \nu'_0 &= \nu_0 + n \\ \nu'_0 \sigma_0'^2 &= \nu_0 \sigma_0^2 + \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n_0 n}{n_0 + n} (\mu_0 - \bar{x})^2\end{aligned}$$

The respective numbers of pseudo-observations add the number of actual observations to them. The new mean hyperparameter is once again a weighted average, this time weighted by the relative numbers of observations. Finally, the update for $\nu'_0 \sigma_0'^2$ is similar to the case with known mean, but in this case the sum of squared deviations is taken with respect to the observed data mean rather than the true mean, and as a result a new "interaction term" needs to be added to take care of the additional error source stemming from the deviation between prior and data mean.

Proof is as follows.

Occurrence

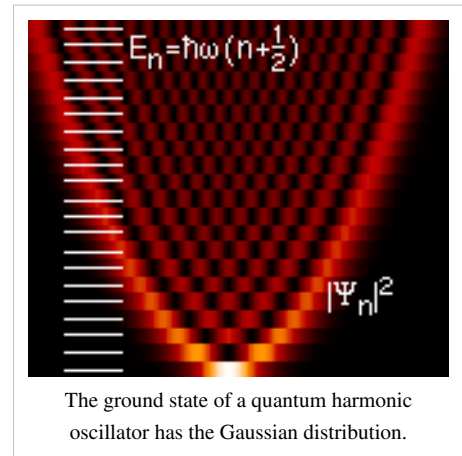
The occurrence of normal distribution in practical problems can be loosely classified into three categories:

1. Exactly normal distributions;
2. Approximately normal laws, for example when such approximation is justified by the central limit theorem; and
3. Distributions modeled as normal – the normal distribution being the distribution with maximum entropy for a given mean and variance.

Exact normality

Certain quantities in physics are distributed normally, as was first demonstrated by James Clerk Maxwell. Examples of such quantities are:

- Velocities of the molecules in the ideal gas. More generally, velocities of the particles in any system in thermodynamic equilibrium will have normal distribution, due to the maximum entropy principle.
- Probability density function of a ground state in a quantum harmonic oscillator.
- The position of a particle that experiences diffusion. If initially the particle is located at a specific point (that is its probability distribution is the dirac delta function), then after time t its location is described by a normal distribution with variance t , which satisfies the diffusion equation $\partial/\partial t f(x,t) = 1/2 \partial^2/\partial x^2 f(x,t)$. If the initial location is given by a certain density function $g(x)$, then the density at time t is the convolution of g and the normal PDF.



The ground state of a quantum harmonic oscillator has the Gaussian distribution.

Approximate normality

Approximately normal distributions occur in many situations, as explained by the central limit theorem. When the outcome is produced by many small effects acting *additively and independently*, its distribution will be close to normal. The normal approximation will not be valid if the effects act multiplicatively (instead of additively), or if there is a single external influence that has a considerably larger magnitude than the rest of the effects.

- In counting problems, where the central limit theorem includes a discrete-to-continuum approximation and where infinitely divisible and decomposable distributions are involved, such as
 - Binomial random variables, associated with binary response variables;
 - Poisson random variables, associated with rare events;
- Thermal light has a Bose–Einstein distribution on very short time scales, and a normal distribution on longer timescales due to the central limit theorem.

Assumed normality

I can only recognize the occurrence of the normal curve – the Laplacian curve of errors – as a very abnormal phenomenon. It is roughly approximated to in certain distributions; for this reason, and on account for its beautiful simplicity, we may, perhaps, use it as a first approximation, particularly in theoretical investigations.

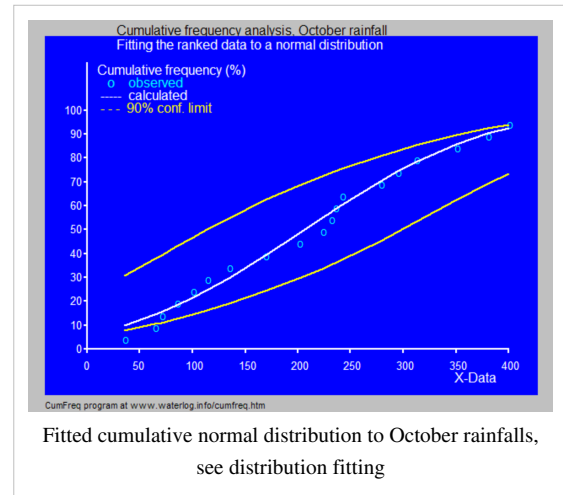
—Pearson (1901)

There are statistical methods to empirically test that assumption, see the above Normality tests section.

- In biology, the *logarithm* of various variables tend to have a normal distribution, that is, they tend to have a log-normal distribution (after separation on male/female subpopulations), with examples including:
 - Measures of size of living tissue (length, height, skin area, weight);
 - The *length* of *inert* appendages (hair, claws, nails, teeth) of biological specimens, *in the direction of growth*; presumably the thickness of tree bark also falls under this category;
 - Certain physiological measurements, such as blood pressure of adult humans.
- In finance, in particular the Black–Scholes model, changes in the *logarithm* of exchange rates, price indices, and stock market indices are assumed normal (these variables behave like compound interest, not like simple interest, and so are multiplicative). Some mathematicians such as Benoît Mandelbrot have argued that log-Levy distributions, which possesses heavy tails would be a more appropriate model, in particular for the analysis for stock market crashes.
- Measurement errors in physical experiments are often modeled by a normal distribution. This use of a normal distribution does not imply that one is assuming the measurement errors are normally distributed, rather using the normal distribution produces the most conservative predictions possible given only knowledge about the mean and variance of the errors.



- In standardized testing, results can be made to have a normal distribution by either selecting the number and difficulty of questions (as in the IQ test) or transforming the raw test scores into "output" scores by fitting them to the normal distribution. For example, the SAT's traditional range of 200–800 is based on a normal distribution with a mean of 500 and a standard deviation of 100.
- Many scores are derived from the normal distribution, including percentile ranks ("percentiles" or "quantiles"), normal curve equivalents, stanines, z-scores, and T-scores. Additionally, some behavioral statistical procedures assume that scores are normally distributed; for example, t-tests and ANOVAs. Bell curve grading assigns relative grades based on a normal distribution of scores.
- In hydrology the distribution of long duration river discharge or rainfall, e.g. monthly and yearly totals, is often thought to be practically normal according to the central limit theorem. The blue picture illustrates an example of fitting the normal distribution to ranked October rainfalls showing the 90% confidence belt based on the binomial distribution. The rainfall data are represented by plotting positions as part of the cumulative frequency analysis.

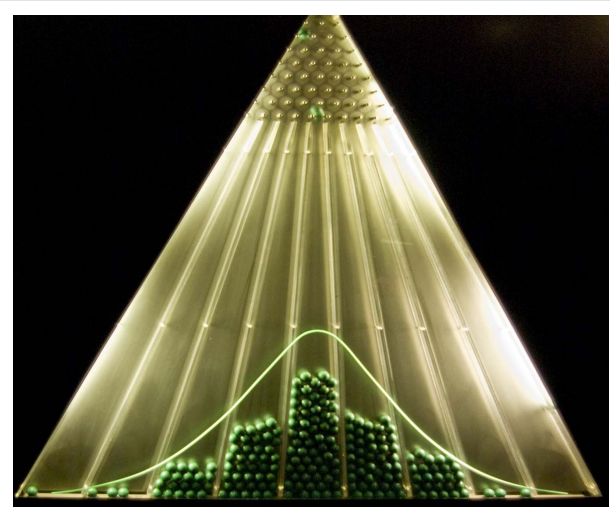


Generating values from normal distribution

In computer simulations, especially in applications of the Monte-Carlo method, it is often desirable to generate values that are normally distributed. The algorithms listed below all generate the standard normal deviates, since a $N(\mu, \sigma^2)$

can be generated as $X = \mu + \sigma Z$, where Z is standard normal. All these algorithms rely on the availability of a random number generator U capable of producing uniform random variates.

- The most straightforward method is based on the probability integral transform property: if U is distributed uniformly on $(0,1)$, then $\Phi^{-1}(U)$ will have the standard normal distribution. The drawback of this method is that it relies on calculation of the probit function Φ^{-1} , which cannot be done analytically. Some approximate methods are described in Hart (1968) and in the erf article. Wichura gives a fast algorithm for computing this function to 16 decimal places, which is used by R to compute random variates of the normal distribution.
- An easy to program approximate approach, that relies on the central limit theorem, is as follows: generate 12 uniform $U(0,1)$ deviates, add them all up, and subtract 6 – the resulting random variable will have approximately standard normal distribution. In truth, the distribution will be Irwin–Hall, which is a 12-section eleventh-order polynomial approximation to the normal distribution. This random deviate will have a limited range of $(-6, 6)$.



The bean machine, a device invented by Francis Galton, can be called the first generator of normal random variables. This machine consists of a vertical board with interleaved rows of pins. Small balls are dropped from the top and then bounce randomly left or right as they hit the pins. The balls are collected into bins at the bottom and settle down into a pattern resembling the Gaussian curve.

- The Box–Muller method uses two independent random numbers U and V distributed uniformly on $(0,1)$. Then the two random variables X and Y

$$X = \sqrt{-2 \ln U} \cos(2\pi V),$$

$$Y = \sqrt{-2 \ln U} \sin(2\pi V).$$

will both have the standard normal distribution, and will be independent. This formulation arises because for a bivariate normal random vector $(X Y)$ the squared norm $X^2 + Y^2$ will have the chi-squared distribution with two degrees of freedom, which is an easily generated exponential random variable corresponding to the quantity $-2\ln(U)$ in these equations; and the angle is distributed uniformly around the circle, chosen by the random variable V .

- Marsaglia polar method is a modification of the Box–Muller method algorithm, which does not require computation of functions $\sin()$ and $\cos()$. In this method U and V are drawn from the uniform $(-1,1)$ distribution, and then $S = U^2 + V^2$ is computed. If S is greater or equal to one then the method starts over, otherwise two quantities

$$X = U \sqrt{\frac{-2 \ln S}{S}}, \quad Y = V \sqrt{\frac{-2 \ln S}{S}}$$

are returned. Again, X and Y will be independent and standard normally distributed.

- The Ratio method is a rejection method. The algorithm proceeds as follows:
 - Generate two independent uniform deviates U and V ;
 - Compute $X = \sqrt{8/e} (V - 0.5)/U$;
 - Optional: if $X^2 \leq 5 - 4e^{1/4}U$ then accept X and terminate algorithm;
 - Optional: if $X^2 \geq 4e^{-1.35}/U + 1.4$ then reject X and start over from step 1;
 - If $X^2 \leq -4 \ln U$ then accept X , otherwise start over the algorithm.
- The ziggurat algorithm is faster than the Box–Muller transform and still exact. In about 97% of all cases it uses only two random numbers, one random integer and one random uniform, one multiplication and an if-test. Only in 3% of the cases, where the combination of those two falls outside the "core of the ziggurat" (a kind of rejection sampling using logarithms), do exponentials and more uniform random numbers have to be employed.
- There is also some investigation into the connection between the fast Hadamard transform and the normal distribution, since the transform employs just addition and subtraction and by the central limit theorem random numbers from almost any distribution will be transformed into the normal distribution. In this regard a series of Hadamard transforms can be combined with random permutations to turn arbitrary data sets into a normally distributed data.

Numerical approximations for the normal CDF

The standard normal CDF is widely used in scientific and statistical computing. The values $\Phi(x)$ may be approximated very accurately by a variety of methods, such as numerical integration, Taylor series, asymptotic series and continued fractions. Different approximations are used depending on the desired level of accuracy.

- Zelen & Severo (1964) give the approximation for $\Phi(x)$ for $x > 0$ with the absolute error $|\varepsilon(x)| < 7.5 \cdot 10^{-8}$ (algorithm 26.2.17^[10]):

$$\Phi(x) = 1 - \phi(x) \left(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5 \right) + \varepsilon(x), \quad t = \frac{1}{1 + b_0 x},$$

where $\phi(x)$ is the standard normal PDF, and $b_0 = 0.2316419$, $b_1 = 0.319381530$, $b_2 = -0.356563782$, $b_3 = 1.781477937$, $b_4 = -1.821255978$, $b_5 = 1.330274429$.

- Hart (1968) lists almost a hundred of rational function approximations for the $\text{erfc}()$ function. His algorithms vary in the degree of complexity and the resulting precision, with maximum absolute precision of 24 digits. An algorithm by West (2009) combines Hart's algorithm 5666 with a continued fraction approximation in the tail to

provide a fast computation algorithm with a 16-digit precision.

- Cody (1969) after recalling Hart68 solution is not suited for erf , gives a solution for both erf and $erfc$, with maximal relative error bound, via Rational Chebyshev Approximation.
- Marsaglia (2004) suggested a simple algorithm^[11] based on the Taylor series expansion

$$\Phi(x) = \frac{1}{2} + \phi(x) \left(x + \frac{x^3}{3} + \frac{x^5}{3 \cdot 5} + \frac{x^7}{3 \cdot 5 \cdot 7} + \frac{x^9}{3 \cdot 5 \cdot 7 \cdot 9} + \dots \right)$$

for calculating $\Phi(x)$ with arbitrary precision. The drawback of this algorithm is comparatively slow calculation time (for example it takes over 300 iterations to calculate the function with 16 digits of precision when $x = 10$).

- The GNU Scientific Library calculates values of the standard normal CDF using Hart's algorithms and approximations with Chebyshev polynomials.

History

Development

Some authors attribute the credit for the discovery of the normal distribution to de Moivre, who in 1738^[12] published in the second edition of his "*The Doctrine of Chances*" the study of the coefficients in the binomial expansion of $(a + b)^n$. De Moivre proved that the middle term in this expansion has the approximate magnitude of $2/\sqrt{2\pi n}$, and that "If m or $\frac{1}{2}n$ be a Quantity infinitely great, then the Logarithm of the Ratio, which a Term distant from the middle by the Interval \square , has to the middle Term, is $-\frac{2\square^2}{n}$."^[13] Although this theorem can be interpreted as the first obscure expression for the normal probability law, Stigler points out that de Moivre himself did not interpret his results as anything more than the approximate rule for the binomial coefficients, and in particular de Moivre lacked the concept of the probability density function.



Carl Friedrich Gauss discovered the normal distribution in 1809 as a way to rationalize the method of least squares.

In 1809 Gauss published his monograph "*Theoria motus corporum coelestium in sectionibus conicis solem ambientium*" where among other things he introduces several important statistical concepts, such as the method of least squares, the method of maximum likelihood, and the *normal distribution*. Gauss used M, M', M'', \dots to denote the measurements of some unknown quantity V , and sought the "most probable" estimator: the one that maximizes the probability $\varphi(M-V) \cdot \varphi(M'-V) \cdot \varphi(M''-V) \cdot \dots$ of obtaining the observed experimental results. In his notation $\varphi\Delta$ is the probability law of the measurement errors of magnitude Δ . Not knowing what the function φ is, Gauss requires that his method should reduce to the well-known answer: the arithmetic mean of the measured values.^[14] Starting from these principles, Gauss demonstrates that the only law that rationalizes the choice of arithmetic mean as an estimator of the location parameter, is the normal law of errors:

$$\varphi\Delta = \frac{h}{\sqrt{\pi}} e^{-hh\Delta\Delta},$$

where h is "the measure of the precision of the observations". Using this normal law as a generic model for errors in the experiments, Gauss formulates what is now known as the non-linear weighted least squares (NWLS) method.

Although Gauss was the first to suggest the normal distribution law, Laplace made significant contributions.^[15] It was Laplace who first posed the problem of aggregating several observations in 1774, although his own solution led to the Laplacian distribution. It was Laplace who first calculated the value of the integral $\int e^{-t^2} dt = \sqrt{\pi}$ in 1782, providing the normalization constant for the normal distribution. Finally, it was Laplace who in 1810 proved and presented to the Academy the fundamental central limit theorem, which emphasized the theoretical importance of the normal distribution.

It is of interest to note that in 1809 an American mathematician Adrain published two derivations of the normal probability law, simultaneously and independently from Gauss. His works remained largely unnoticed by the scientific community, until in 1871 they were "rediscovered" by Abbe.

In the middle of the 19th century Maxwell demonstrated that the normal distribution is not just a convenient mathematical tool, but may also occur in natural phenomena: "The number of particles whose velocity, resolved in a certain direction, lies between x and $x + dx$ is

$$N \frac{1}{\alpha \sqrt{\pi}} e^{-\frac{x^2}{\alpha^2}} dx$$

Naming

Since its introduction, the normal distribution has been known by many different names: the law of error, the law of facility of errors, Laplace's second law, Gaussian law, etc. Gauss himself apparently coined the term with reference to the "normal equations" involved in its applications, with normal having its technical meaning of orthogonal rather than "usual".^[16] However, by the end of the 19th century some authors^[17] had started using the name *normal distribution*, where the word "normal" was used as an adjective – the term now being seen as a reflection of the fact that this distribution was seen as typical, common – and thus "normal". Peirce (one of those authors) once defined "normal" thus: "...the 'normal' is not the average (or any other kind of mean) of what actually occurs, but of what *would*, in the long run, occur under certain circumstances."^[18] Around the turn of the 20th century Pearson popularized the term *normal* as a designation for this distribution.

Many years ago I called the Laplace–Gaussian curve the *normal* curve, which name, while it avoids an international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another 'abnormal'.

—Pearson (1920)

Also, it was Pearson who first wrote the distribution in terms of the standard deviation σ as in modern notation. Soon after this, in year 1915, Fisher added the location parameter to the formula for normal distribution, expressing it in the way it is written nowadays:

$$df = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

The term "standard normal", which denotes the normal distribution with zero mean and unit variance came into general use around 1950s, appearing in the popular textbooks by P.G. Hoel (1947) "*Introduction to mathematical statistics*" and A.M. Mood (1950) "*Introduction to the theory of statistics*".

When the name is used, the "Gaussian distribution" was named after Carl Friedrich Gauss, who introduced the distribution in 1809 as a way of rationalizing the method of least squares as outlined above. Among English



Marquis de Laplace proved the central limit theorem in 1810, consolidating the importance of the normal distribution in statistics.

speakers, both "normal distribution" and "Gaussian distribution" are in common use, with different terms preferred by different communities.

Notes

- [1] *Normal Distribution* (http://findarticles.com/p/articles/mi_g2699/is_0002/ai_2699000241), Gale Encyclopedia of Psychology
- [2] WolframAlpha.com ([http://www.wolframalpha.com/input/?i=Table\[{N\(Erf\(n/Sqrt\(2\)\),+12\),+N\(1-Erf\(n/Sqrt\(2\)\),+12\),+N\(1/\(1-Erf\(n/Sqrt\(2\)\)\),+12\)},+{n,1,6}\]](http://www.wolframalpha.com/input/?i=Table[{N(Erf(n/Sqrt(2)),+12),+N(1-Erf(n/Sqrt(2)),+12),+N(1/(1-Erf(n/Sqrt(2))),+12)},+{n,1,6}]))
- [3] part 1 ([http://www.wolframalpha.com/input/?i=Table\[Sqrt\(2\)*InverseErf\(x\),+{x,+N\({8/10,+9/10,+19/20,+49/50,+99/100,+995/1000,+998/1000},+13\)}\]](http://www.wolframalpha.com/input/?i=Table[Sqrt(2)*InverseErf(x),+{x,+N({8/10,+9/10,+19/20,+49/50,+99/100,+995/1000,+998/1000},+13)}])), part 2 ([http://www.wolframalpha.com/input/?i=Table\[{N\(1-10^\(-x\),9\),N\(Sqrt\(2\)*InverseErf\(1-10^\(-x\)\),13\)},+{x,3,9}\]](http://www.wolframalpha.com/input/?i=Table[{N(1-10^(-x),9),N(Sqrt(2)*InverseErf(1-10^(-x)),13)},+{x,3,9}]))
- [4] Normal Approximation to Poisson(λ) Distribution, <http://www.stat.ucla.edu/> (http://www.stat.ucla.edu/~dinov/courses_students.dir/Applets.dir/NormalApprox2PoissonApplet.html)
- [5] Quine, M.P. (1993) "On three characterisations of the normal distribution" (<http://www.math.uni.wroc.pl/~pms/publicationsArticle.php?nr=14.2&nrA=8&ppB=257&ppE=263>), *Probability and Mathematical Statistics*, 14 (2), 257-263
- [6] UIUC, Lecture 21. *The Multivariate Normal Distribution* (<http://www.math.uiuc.edu/~r-ash/Stat/StatLec21-25.pdf>), 21.6: "Individually Gaussian Versus Jointly Gaussian".
- [7] Edward L. Melnick and Aaron Tenenbein, "Misspecifications of the Normal Distribution", *The American Statistician*, volume 36, number 4 November 1982, pages 372–373
- [8] <http://www.allisons.org/ll/MML/KL/Normal/>
- [9] *Normal Product Distribution* (<http://mathworld.wolfram.com/NormalProductDistribution.html>), Mathworld
- [10] http://www.math.sfu.ca/~cbm/aands/page_932.htm
- [11] For example, this algorithm is given in the article [Bc programming language](#).
- [12] De Moivre first published his findings in 1733, in a pamphlet "Approximatio ad Summam Terminorum Binomiali in Seriem Expansi" that was designated for private circulation only. But it was not until the year 1738 that he made his results publicly available. The original pamphlet was reprinted several times, see for example .
- [13] De Moivre, Abraham (1733), Corollary I – see
- [14] "It has been customary certainly to regard as an axiom the hypothesis that if any quantity has been determined by several direct observations, made under the same circumstances and with equal care, the arithmetical mean of the observed values affords the most probable value, if not rigorously, yet very nearly at least, so that it is always most safe to adhere to it." —
- [15] "My custom of terming the curve the Gauss–Laplacian or *normal* curve saves us from proportioning the merit of discovery between the two great astronomer mathematicians." quote from
- [16] Jaynes, Edwin J.; *Probability Theory: The Logic of Science*, Ch 7 (<http://www.biba.inrialpes.fr/Jaynes/cc07s.pdf>)
- [17] Besides those specifically referenced here, such use is encountered in the works of Peirce, Galton () and Lexis (,) c. 1875.
- [18] Peirce, Charles S. (c. 1909 MS), *Collected Papers* v. 6, paragraph 327

Citations

References

- Aldrich, John; Miller, Jeff. "Earliest Uses of Symbols in Probability and Statistics" (<http://jeff560.tripod.com/stat.html>).
- Aldrich, John; Miller, Jeff. "Earliest Known Uses of Some of the Words of Mathematics" (<http://jeff560.tripod.com/mathword.html>). In particular, the entries for "bell-shaped and bell curve" (<http://jeff560.tripod.com/b.html>), "normal (distribution)" (<http://jeff560.tripod.com/n.html>), "Gaussian" (<http://jeff560.tripod.com/g.html>), and "Error, law of error, theory of errors, etc." (<http://jeff560.tripod.com/e.html>).
- Amari, Shun-ichi; Nagaoka, Hiroshi (2000). *Methods of Information Geometry*. Oxford University Press. ISBN 0-8218-0531-2.
- Bernardo, José M.; Smith, Adrian F. M. (2000). *Bayesian Theory*. Wiley. ISBN 0-471-49464-X.
- Bryc, Włodzimierz (1995). *The Normal Distribution: Characterizations with Applications*. Springer-Verlag. ISBN 0-387-97990-5.
- Casella, George; Berger, Roger L. (2001). *Statistical Inference* (2nd ed.). Duxbury. ISBN 0-534-24312-6.
- Cody, William J. (1969). "Rational Chebyshev Approximations for the Error Function". *Mathematics of Computation* **23** (107): 631–638. doi: 10.1090/S0025-5718-1969-0247736-4 (<http://dx.doi.org/10.1090>)

- S0025-5718-1969-0247736-4).
- Cover, Thomas M.; Thomas, Joy A. (2006). *Elements of Information Theory*. John Wiley and Sons.
 - de Moivre, Abraham (1738). *The Doctrine of Chances*. ISBN 0-8218-2103-2.
 - Fan, Jianqing (1991). "On the optimal rates of convergence for nonparametric deconvolution problems". *The Annals of Statistics* **19** (3): 1257–1272. doi: 10.1214/aos/1176348248 (<http://dx.doi.org/10.1214/aos/1176348248>). JSTOR 2241949 (<http://www.jstor.org/stable/2241949>).
 - Galton, Francis (1889). *Natural Inheritance* (<http://galton.org/books/natural-inheritance/pdf/galton-nat-inh-lup-clean.pdf>). London, UK: Richard Clay and Sons.
 - Galambos, Janos; Simonelli, Italo (2004). *Products of Random Variables: Applications to Problems of Physics and to Arithmetical Functions*. Marcel Dekker, Inc. ISBN 0-8247-5402-6.
 - Gauss, Carolo Friderico (1809). *Theoria motvs corporvm coelestivm in sectionibvs conicis Solem ambientivm* [*Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*] (in Latin). English translation (<http://books.google.com/books?id=1TIAAAAAQAAJ>).
 - Gould, Stephen Jay (1981). *The Mismeasure of Man* (first ed.). W. W. Norton. ISBN 0-393-01489-4.
 - Halperin, Max; Hartley, Herman O.; Hoel, Paul G. (1965). "Recommended Standards for Statistical Symbols and Notation. COPSS Committee on Symbols and Notation". *The American Statistician* **19** (3): 12–14. doi: 10.2307/2681417 (<http://dx.doi.org/10.2307/2681417>). JSTOR 2681417 (<http://www.jstor.org/stable/2681417>).
 - Hart, John F.; et al. (1968). *Computer Approximations*. New York, NY: John Wiley & Sons, Inc. ISBN 0-88275-642-7.
 - Hazewinkel, Michiel, ed. (2001), "Normal Distribution" (<http://www.encyclopediaofmath.org/index.php?title=p/n067460>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
 - Herrnstein, Richard J.; Murray, Charles (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press. ISBN 0-02-914673-9.
 - Huxley, Julian S. (1932). *Problems of Relative Growth*. London. ISBN 0-486-61114-0. OCLC 476909537 (<http://www.worldcat.org/oclc/476909537>).
 - Johnson, Norman L.; Kotz, Samuel; Balakrishnan, Narayanaswamy (1994). *Continuous Univariate Distributions, Volume 1*. Wiley. ISBN 0-471-58495-9.
 - Johnson, Norman L.; Kotz, Samuel; Balakrishnan, Narayanaswamy (1995). *Continuous Univariate Distributions, Volume 2*. Wiley. ISBN 0-471-58494-0.
 - Kinderman, Albert J.; Monahan, John F. (1977). "Computer Generation of Random Variables Using the Ratio of Uniform Deviates". *ACM Transactions on Mathematical Software* **3** (3): 257–260. doi: 10.1145/355744.355750 (<http://dx.doi.org/10.1145/355744.355750>).
 - Krishnamoorthy, Kalimuthu (2006). *Handbook of Statistical Distributions with Applications*. Chapman & Hall/CRC. ISBN 1-58488-635-8.
 - Kruskal, William H.; Stigler, Stephen M. (1997). Spencer, Bruce D., ed. *Normative Terminology: 'Normal' in Statistics and Elsewhere*. Statistics and Public Policy. Oxford University Press. ISBN 0-19-852341-6.
 - Laplace, Pierre-Simon de (1774). "Mémoire sur la probabilité des causes par les événements" (<http://gallica.bnf.fr/ark:/12148/bpt6k77596b/f32>). *Mémoires de l'Académie royale des Sciences de Paris (Savants étrangers)*, tome 6: 621–656. Translated by Stephen M. Stigler in *Statistical Science* **1** (3), 1986: JSTOR 2245476 (<http://www.jstor.org/stable/2245476>).
 - Laplace, Pierre-Simon (1812). *Théorie analytique des probabilités* [*Analytical theory of probabilities*].
 - Le Cam, Lucien; Lo Yang, Grace (2000). *Asymptotics in Statistics: Some Basic Concepts* (second ed.). Springer. ISBN 0-387-95036-2.
 - Lexis, Wilhelm (1878). "Sur la durée normale de la vie humaine et sur la théorie de la stabilité des rapports statistiques". *Annales de démographie internationale* (Paris) **II**: 447–462.

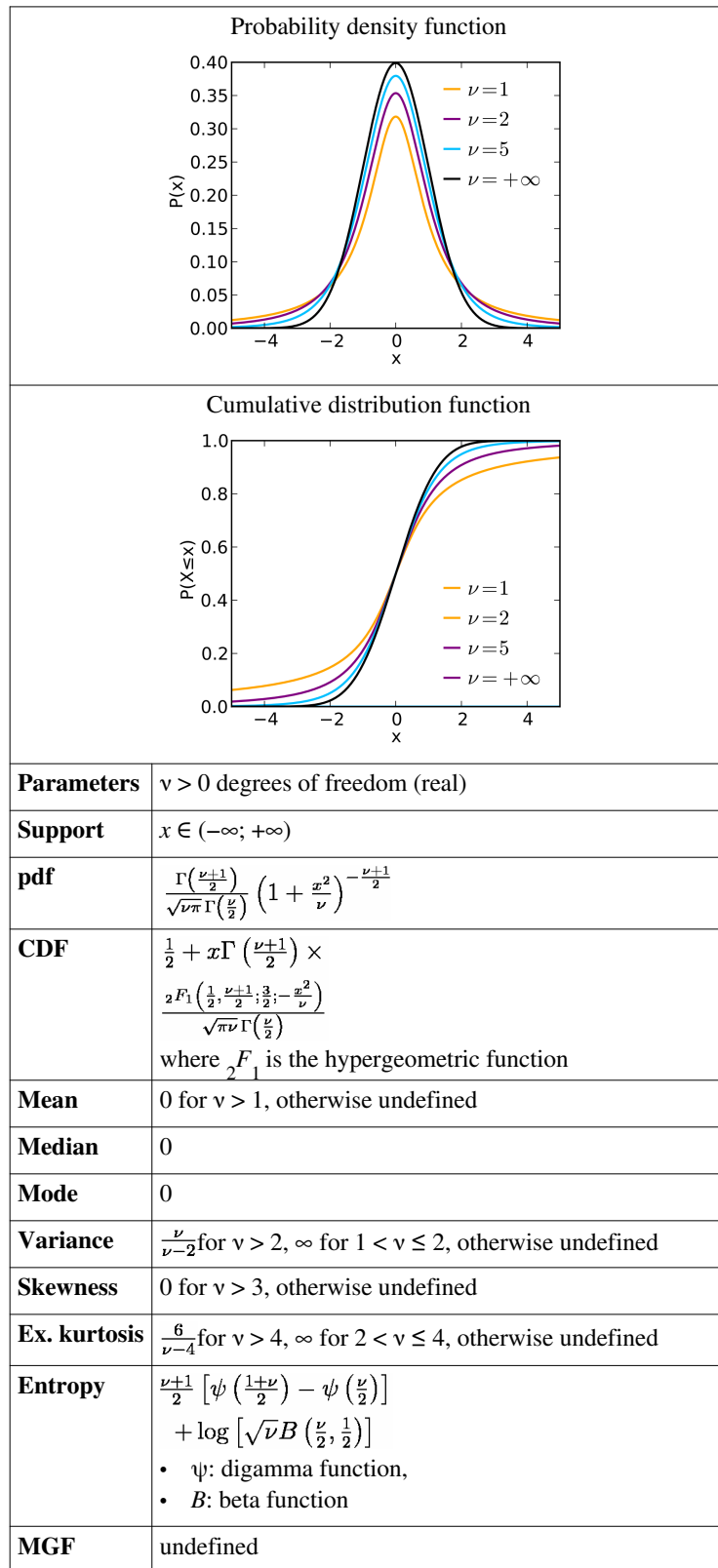
- Lukacs, Eugene; King, Edgar P. (1954). "A Property of Normal Distribution". *The Annals of Mathematical Statistics* **25** (2): 389–394. doi: 10.1214/aoms/1177728796 (<http://dx.doi.org/10.1214/aoms/1177728796>). JSTOR 2236741 (<http://www.jstor.org/stable/2236741>).
- McPherson, Glen (1990). *Statistics in Scientific Investigation: Its Basis, Application and Interpretation*. Springer-Verlag. ISBN 0-387-97137-8.
- Marsaglia, George; Tsang, Wai Wan (2000). "The Ziggurat Method for Generating Random Variables" (<http://www.jstatsoft.org/v05/i08/paper>). *Journal of Statistical Software* **5** (8).
- Wallace, C. S. (1996). "Fast pseudo-random generators for normal and exponential variates". *ACM Transactions on Mathematical Software* **22** (1): 119–127. doi: 10.1145/225545.225554 (<http://dx.doi.org/10.1145/225545.225554>).
- Marsaglia, George (2004). "Evaluating the Normal Distribution" (<http://www.jstatsoft.org/v11/i05/paper>). *Journal of Statistical Software* **11** (4).
- Maxwell, James Clerk (1860). "V. Illustrations of the dynamical theory of gases. — Part I: On the motions and collisions of perfectly elastic spheres". *Philosophical Magazine, series 4* **19** (124): 19–32. doi: 10.1080/14786446008642818 (<http://dx.doi.org/10.1080/14786446008642818>).
- Patel, Jagdish K.; Read, Campbell B. (1996). *Handbook of the Normal Distribution* (2nd ed.). CRC Press. ISBN 0-8247-9342-0.
- Pearson, Karl (1905). "'Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson'. A rejoinder". *Biometrika* **4** (1): 169–212. JSTOR 2331536 (<http://www.jstor.org/stable/2331536>).
- Pearson, Karl (1920). "Notes on the History of Correlation". *Biometrika* **13** (1): 25–45. doi: 10.1093/biomet/13.1.25 (<http://dx.doi.org/10.1093/biomet/13.1.25>). JSTOR 2331722 (<http://www.jstor.org/stable/2331722>).
- Rohrbasser, Jean-Marc; Véron, Jacques (2003). "Wilhelm Lexis: The Normal Length of Life as an Expression of the "Nature of Things"" (http://www.persee.fr/web/revues/home/prescript/article/pop_1634-2941_2003_num_58_3_18444). *Population* **58** (3): 303–322.
- Stigler, Stephen M. (1978). "Mathematical Statistics in the Early States". *The Annals of Statistics* **6** (2): 239–265. doi: 10.1214/aos/1176344123 (<http://dx.doi.org/10.1214/aos/1176344123>). JSTOR 2958876 (<http://www.jstor.org/stable/2958876>).
- Stigler, Stephen M. (1982). "A Modest Proposal: A New Standard for the Normal". *The American Statistician* **36** (2): 137–138. doi: 10.2307/2684031 (<http://dx.doi.org/10.2307/2684031>). JSTOR 2684031 (<http://www.jstor.org/stable/2684031>).
- Stigler, Stephen M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press. ISBN 0-674-40340-1.
- Stigler, Stephen M. (1999). *Statistics on the Table*. Harvard University Press. ISBN 0-674-83601-4.
- Walker, Helen M. (1985). "De Moivre on the Law of Normal Probability" (<http://www.york.ac.uk/depts/math/histstat/demoivre.pdf>). In Smith, David Eugene. *A Source Book in Mathematics*. Dover. ISBN 0-486-64690-4.
- Weisstein, Eric W.. "Normal Distribution" (<http://mathworld.wolfram.com/NormalDistribution.html>). MathWorld.
- West, Graeme (2009). "Better Approximations to Cumulative Normal Functions" (http://www.wilmott.com/pdfs/090721_west.pdf). *Wilmott Magazine*: 70–76.
- Zelen, Marvin; Severo, Norman C. (1964). *Probability Functions (chapter 26)* (http://www.math.sfu.ca/~cbm/aands/page_931.htm). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, by Abramowitz, M.; and Stegun, I. A.: National Bureau of Standards. New York, NY: Dover. ISBN 0-486-61272-4.

External links

- Hazewinkel, Michiel, ed. (2001), "Normal distribution" (<http://www.encyclopediaofmath.org/index.php?title=p/n067460>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
 - Normal Distribution Video Tutorial Part 1-2 (http://www.youtube.com/watch?v=kB_kYUbS_ig)
 - An 8-foot-tall (2.4 m) Probability Machine (named Sir Francis) comparing stock market returns to the randomness of the beans dropping through the quincunx pattern. (<http://www.youtube.com/watch?v=AUSKtk9ENzg>) YouTube link originating from Index Funds Advisors (<http://www.ifa.com>)
-

Student's t-distribution

Student's *t*



CF	$\frac{K_{\nu/2}(\sqrt{\nu} t) \cdot (\sqrt{\nu} t)^{\nu/2}}{\Gamma(\nu/2) 2^{\nu/2-1}}$ for $\nu > 0$ • $K_{\nu}(x)$: Modified Bessel function of the second kind ^[1]
-----------	---

In probability and statistics, **Student's *t*-distribution** (or simply the ***t*-distribution**) is a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. It plays a role in a number of widely used statistical analyses, including the Student's *t*-test for assessing the statistical significance of the difference between two sample means, the construction of confidence intervals for the difference between two population means, and in linear regression analysis. The Student's *t*-distribution also arises in the Bayesian analysis of data from a normal family.

If we take a sample of $n = \nu + 1$ observations from a normal distribution (the black curve on the figure on the right of this page, representing a very large ν), compute the sample mean and plot it, and repeat this process infinitely many times (for the same n), we get the probability density function for that n , as shown in the image on the right.

If we also compute the sample variance for these n observations, then the *t*-distribution (for $n - 1$) can be defined as the distribution of the location of the true mean, relative to the sample mean and divided by the sample standard deviation, after multiplying by the normalizing term \sqrt{n} , where n is the sample size. In this way, the *t*-distribution can be used to estimate how likely it is that the true mean lies in any given range.

The *t*-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean. This makes it useful for understanding the statistical behavior of certain types of ratios of random quantities, in which variation in the denominator is amplified and may produce outlying values when the denominator of the ratio falls close to zero. The Student's *t*-distribution is a special case of the generalised hyperbolic distribution.

History and etymology

In statistics, the *t*-distribution was first derived as a posterior distribution in 1876 by Helmert and Lüroth.

In the English-language literature it takes its name from William Sealy Gosset's 1908 paper in *Biometrika* under the pseudonym "Student".^[2] Gosset worked at the Guinness Brewery in Dublin, Ireland, and was interested in the problems of small samples, for example of the chemical properties of barley where sample sizes might be as low as 3. One version of the origin of the pseudonym is that Gosset's employer forbade members of its staff from publishing scientific papers, so he had to hide his identity. Another version is that Guinness did not want their competitors to know that they were using the *t*-test to test the quality of raw material.^[3]

Gosset's paper refers to the distribution as the "frequency distribution of standard deviations of samples drawn from a normal population". It became well-known through the work of Ronald A. Fisher, who called the distribution "Student's distribution" and referred to the value as t .^[4]

Definition

Probability density function

Student's *t*-distribution has the probability density function given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where ν is the number of *degrees of freedom* and Γ is the gamma function. This may also be written as

$$f(t) = \frac{1}{\sqrt{\nu} B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where B is the Beta function.

For ν even,

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} = \frac{(\nu - 1)(\nu - 3) \cdots 5 \cdot 3}{2\sqrt{\nu}(\nu - 2)(\nu - 4) \cdots 4 \cdot 2}.$$

For ν odd,

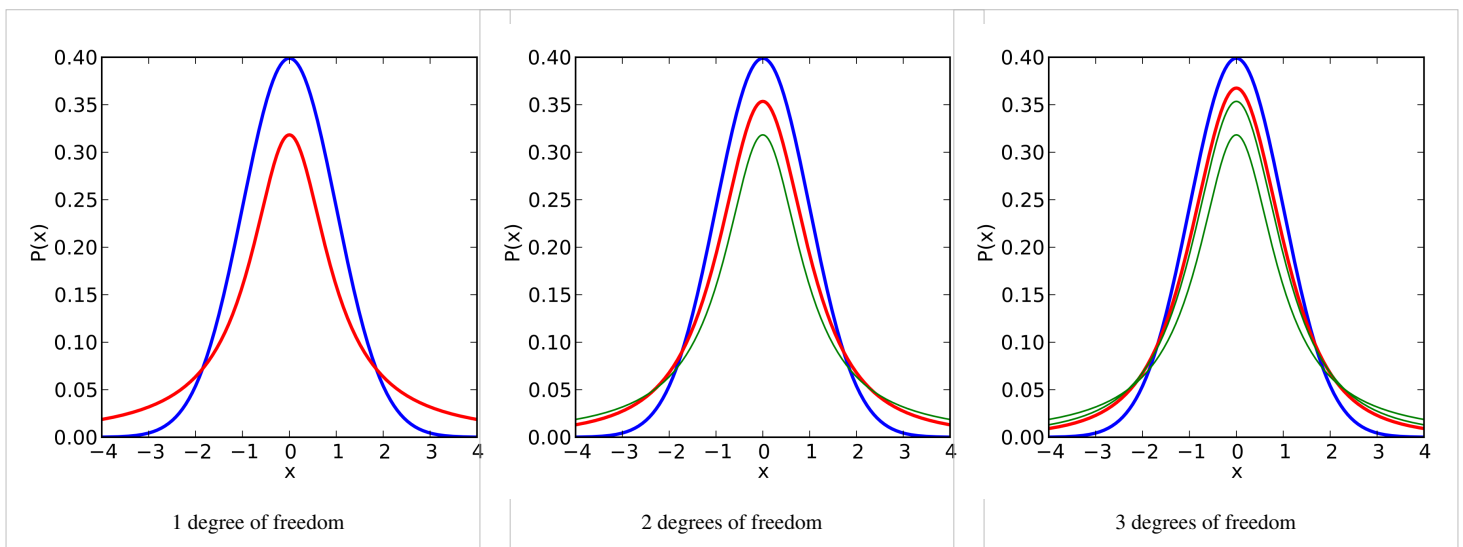
$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} = \frac{(\nu - 1)(\nu - 3) \cdots 4 \cdot 2}{\pi\sqrt{\nu}(\nu - 2)(\nu - 4) \cdots 5 \cdot 3}.$$

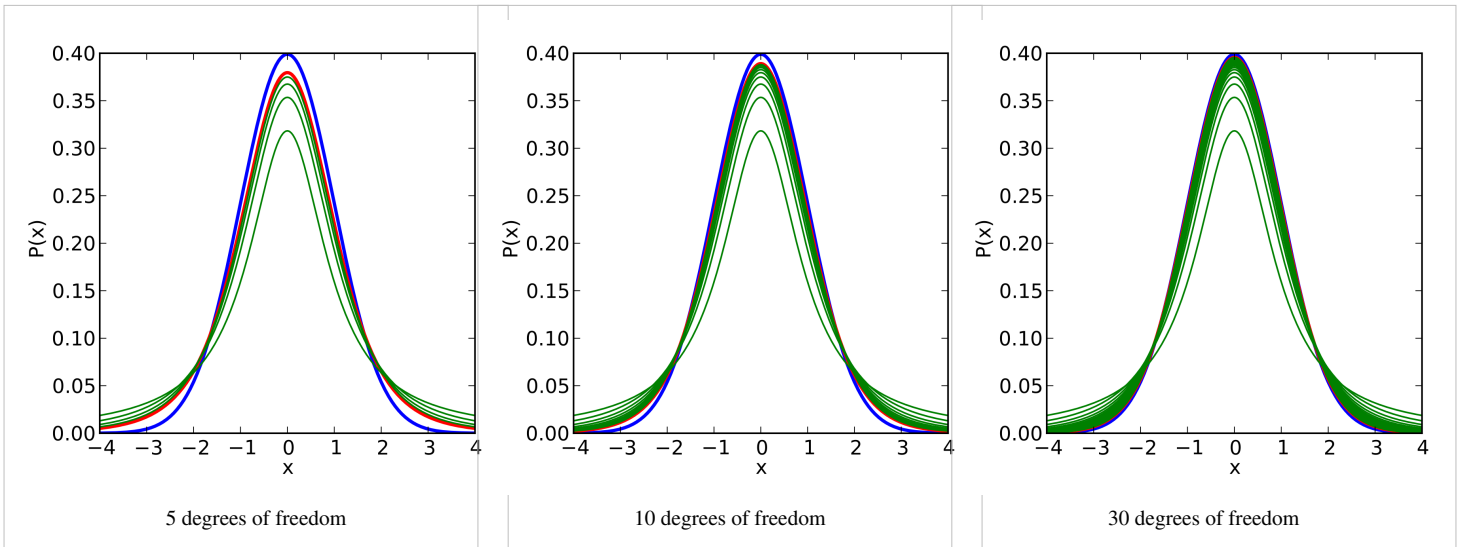
The probability density function is symmetric, and its overall shape resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider. As the number of degrees of freedom grows, the *t*-distribution approaches the normal distribution with mean 0 and variance 1.

The following images show the density of the *t*-distribution for increasing values of ν . The normal distribution is shown as a blue line for comparison. Note that the *t*-distribution (red line) becomes closer to the normal distribution as ν increases.

Density of the *t*-distribution (red) for 1, 2, 3, 5, 10, and 30 degrees of freedom compared to the standard normal distribution (blue).

Previous plots shown in green.





Cumulative distribution function

The cumulative distribution function can be written in terms of I , the regularized incomplete beta function. For $t > 0$,

$$F(t) = \int_{-\infty}^t f(u) du = 1 - \frac{1}{2} I_{x(t)} \left(\frac{\nu}{2}, \frac{1}{2} \right),$$

with

$$x(t) = \frac{\nu}{t^2 + \nu}.$$

Other values would be obtained by symmetry. An alternative formula, valid for $t^2 < \nu$, is

$$\int_{-\infty}^t f(u) du = \frac{1}{2} + t \frac{\Gamma\left(\frac{1}{2}(\nu + 1)\right)}{\sqrt{\pi\nu} \Gamma\left(\frac{\nu}{2}\right)} {}_2F_1\left(\frac{1}{2}, \frac{1}{2}(\nu + 1); \frac{3}{2}; -\frac{t^2}{\nu}\right)$$

where ${}_2F_1$ is a particular case of the hypergeometric function.

Special cases

Certain values of ν give an especially simple form.

- $\nu = 1$

Distribution function:

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan(x).$$

Density function:

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

See Cauchy distribution

- $\nu = 2$

Distribution function:

$$F(x) = \frac{1}{2} + \frac{x}{2\sqrt{2 + x^2}}.$$

Density function:

$$f(x) = \frac{1}{(2 + x^2)^{\frac{3}{2}}}.$$

- $\nu = 3$

Density function:

$$f(x) = \frac{6\sqrt{3}}{\pi (3 + x^2)^2}.$$

- $\nu = \infty$

Density function:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

See Normal distribution

How the *t*-distribution arises

Sampling distribution

Let x_1, \dots, x_n be the numbers observed in a sample from a continuously distributed population with expected value μ . The sample mean and sample variance are given by:

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

The resulting *t*-value is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

The *t*-distribution with $n - 1$ degrees of freedom is the sampling distribution of the *t*-value when the samples consist of independent identically distributed observations from a normally distributed population. Thus for inference purposes *t* is a useful "pivotal quantity" in the case when the mean and variance (μ, σ^2) are unknown population parameters, in the sense that the *t*-value has then a probability distribution that depends on neither μ nor σ^2 .

Bayesian inference

In Bayesian statistics, a (scaled, shifted) *t*-distribution arises as the marginal distribution of the unknown mean of a normal distribution, when the dependence on an unknown variance has been marginalised out:^[5]

$$p(\mu|D, I) = \int p(\mu, \sigma^2|D, I) d\sigma^2$$

$$= \int p(\mu|D, \sigma^2, I) p(\sigma^2|D, I) d\sigma^2$$

where *D* stands for the data $\{x_i\}$ and *I* represents any other information that may have been used to create the model. The distribution is thus the compounding of the conditional distribution of μ given the data and σ^2 with the marginal distribution of σ^2 given the data.

With *n* data points, if uninformative location and scale priors $p(\mu|\sigma^2, I) = \text{const}$ and $p(\sigma^2|I) \propto 1/\sigma^2$ can be taken for μ and σ^2 , then Bayes' theorem gives

$$p(\mu|D, \sigma^2, I) \sim N(\bar{x}, \sigma^2/n)$$

$$p(\sigma^2|D, I) \sim \text{Scale-inv-}\chi^2(\nu, s^2)$$

a Normal distribution and a scaled inverse chi-squared distribution respectively, where $\nu = n - 1$ and

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n - 1}.$$

The marginalisation integral thus becomes

$$\begin{aligned} p(\mu|D, I) &\propto \int_0^\infty \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}n(\mu - \bar{x})^2\right) \cdot \sigma^{-\nu-2} \exp(-\nu s^2/2\sigma^2) d\sigma^2 \\ &\propto \int_0^\infty \sigma^{-\nu-3} \exp\left(-\frac{1}{2\sigma^2}(n(\mu - \bar{x})^2 + \nu s^2)\right) d\sigma^2 \end{aligned}$$

This can be evaluated by substituting $z=A/2\sigma^2$, where $A=n(\mu-\bar{x})^2+\nu s^2$, giving

$$dz = -\frac{A}{2\sigma^4}d\sigma^2,$$

so

$$p(\mu|D, I) \propto A^{-\frac{\nu+1}{2}} \int_0^\infty z^{(\nu-1)/2} \exp(-z) dz$$

But the z integral is now a standard Gamma integral, which evaluates to a constant, leaving

$$\begin{aligned} p(\mu|D, I) &\propto A^{-\frac{\nu+1}{2}} \\ &\propto \left(1 + \frac{n(\mu - \bar{x})^2}{\nu s^2}\right)^{-\frac{\nu+1}{2}} \end{aligned}$$

This is a form of the t distribution with an explicit scaling and shifting that will be explored in more detail in a further section below. It can be related to the standardised t distribution by the substitution

$$t = \frac{\mu - \bar{x}}{s/\sqrt{n}}$$

The derivation above has been presented for the case of uninformative priors for μ and σ^2 ; but it will be apparent that any priors which lead to a Normal distribution being compounded with a scaled inverse chi-squared distribution will lead to a t distribution with scaling and shifting for $P(\mu|D, I)$, although the scaling parameter corresponding to s^2/n above will then be influenced both by the prior information and the data, rather than just by the data as above.

Characterization

As the distribution of a test statistic

Student's t -distribution with ν degrees of freedom can be defined as the distribution of the random variable T with

$$T = \frac{Z}{\sqrt{V/\nu}} = Z\sqrt{\frac{\nu}{V}},$$

where

- Z is normally distributed with expected value 0 and variance 1;
- V has a chi-squared distribution with ν degrees of freedom;
- Z and V are independent.

A different distribution is defined as that of the random variable defined, for a given constant μ , by

$$(Z + \mu)\sqrt{\frac{\nu}{V}}.$$

This random variable has a noncentral t -distribution with noncentrality parameter μ . This distribution is important in studies of the power of Student's t -test.

Derivation

Suppose X_1, \dots, X_n are independent random variables that are normally distributed with expected value μ and variance σ^2 . Let

$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$$

be the sample mean, and

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

be an unbiased estimate of the variance from the sample. It can be shown that the random variable

$$V = (n-1) \frac{S_n^2}{\sigma^2}$$

has a chi-squared distribution with $\nu=n-1$ degrees of freedom (by Cochran's theorem). It is readily shown that the quantity

$$Z = (\bar{X}_n - \mu) \frac{\sqrt{n}}{\sigma}$$

is normally distributed with mean 0 and variance 1, since the sample mean \bar{X}_n is normally distributed with mean μ and variance σ^2/n . Moreover, it is possible to show that these two random variables (the normally distributed one Z and the chi-squared-distributed one V) are independent. Consequently Wikipedia:Please clarify the pivotal quantity,

$$T \equiv \frac{Z}{\sqrt{V/\nu}} = (\bar{X}_n - \mu) \frac{\sqrt{n}}{S_n},$$

which differs from Z in that the exact standard deviation σ is replaced by the random variable S_n , has a Student's t -distribution as defined above. Notice that the unknown population variance σ^2 does not appear in T , since it was in both the numerator and the denominator, so it canceled. Gosset intuitively obtained the probability density function stated above, with ν equal to $n - 1$, and Fisher proved it in 1925.

The distribution of the test statistic, T , depends on ν , but not μ or σ ; the lack of dependence on μ and σ is what makes the t -distribution important in both theory and practice.

As a maximum entropy distribution

Student's t -distribution is the maximum entropy probability distribution for a random variate X for which $E(\ln(\nu + X^2))$ is fixed.

Properties

Moments

The raw moments of the t -distribution are

$$E(T^k) = \begin{cases} 0 & k \text{ odd, } 0 < k < \nu \\ \frac{1}{\sqrt{\pi}\Gamma(\frac{\nu}{2})} \left[\Gamma\left(\frac{k+1}{2}\right) \Gamma\left(\frac{\nu-k}{2}\right) \nu^{\frac{k}{2}} \right] & k \text{ even, } 0 < k < \nu \end{cases}$$

Moments of order ν or higher do not exist.^[7]

The term for $0 < k < \nu, k$ even, may be simplified using the properties of the gamma function to

$$E(T^k) = \nu^{\frac{k}{2}} \prod_{i=1}^{\frac{k}{2}} \frac{2i-1}{\nu-2i} \quad k \text{ even, } 0 < k < \nu.$$

For a t -distribution with ν degrees of freedom, the expected value is 0, and its variance is $\nu/(\nu - 2)$ if $\nu > 2$. The skewness is 0 if $\nu > 3$ and the excess kurtosis is $6/(\nu - 4)$ if $\nu > 4$.

Relation to F distribution

- $Y \sim F(\nu_1 = 1, \nu_2 = \nu)$ has an F -distribution if $Y = X^2$ and $X \sim t(\nu)$ has a Student's t -distribution.

Monte Carlo sampling

There are various approaches to constructing random samples from the Student's t -distribution. The matter depends on whether the samples are required on a stand-alone basis, or are to be constructed by application of a quantile function to uniform samples; e.g., in the multi-dimensional applications basis of copula-dependency.^[citation needed] In the case of stand-alone sampling, an extension of the Box–Muller method and its polar form is easily deployed. It has the merit that it applies equally well to all real positive degrees of freedom, ν , while many other candidate methods fail if ν is close to zero.

Integral of Student's probability density function and p -value

The function $A(t|\nu)$ is the integral of Student's probability density function, $f(t)$ between $-t$ and t , for $t \geq 0$. It thus gives the probability that a value of t less than that calculated from observed data would occur by chance. Therefore, the function $A(t|\nu)$ can be used when testing whether the difference between the means of two sets of data is statistically significant, by calculating the corresponding value of t and the probability of its occurrence if the two sets of data were drawn from the same population. This is used in a variety of situations, particularly in t -tests. For the statistic t , with ν degrees of freedom, $A(t|\nu)$ is the probability that t would be less than the observed value if the two means were the same (provided that the smaller mean is subtracted from the larger, so that $t \geq 0$). It can be easily calculated from the cumulative distribution function $F_\nu(t)$ of the t -distribution:

$$A(t|\nu) = F_\nu(t) - F_\nu(-t) = 1 - I_{\frac{\nu}{\nu+t^2}}\left(\frac{\nu}{2}, \frac{1}{2}\right),$$

where I_x is the regularized incomplete beta function (a, b).

For statistical hypothesis testing this function is used to construct the p -value.

Non-standardized Student's t -distribution

In terms of scaling parameter σ , or σ^2

Student's t distribution can be generalized to a three parameter location-scale family, introducing a location parameter μ and a scale parameter σ , through the relation

$$X = \mu + \sigma T$$

The resulting **non-standardized Student's t -distribution** has a density defined by

$$p(x|\nu, \mu, \sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \left(1 + \frac{1}{\nu} \left(\frac{x - \mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}$$

Here, σ does *not* correspond to a standard deviation: it is not the standard deviation of the scaled t distribution, which may not even exist; nor is it the standard deviation of the underlying normal distribution, which is unknown. σ simply sets the overall scaling of the distribution. In the Bayesian derivation of the marginal distribution of an unknown Normal mean μ above, σ as used here corresponds to the quantity s/\sqrt{n} , where

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n - 1}.$$

Equivalently, the distribution can be written in terms of σ^2 , the square of this scale parameter:

$$p(x|\nu, \mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{1}{\nu} \frac{(x - \mu)^2}{\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

Other properties of this version of the distribution are:

$$\begin{aligned} E(X) &= \mu && \text{for } \nu > 1, \\ \text{var}(X) &= \sigma^2 \frac{\nu}{\nu - 2} && \text{for } \nu > 2, \\ \text{mode}(X) &= \mu. \end{aligned}$$

This distribution results from compounding a Gaussian distribution (normal distribution) with mean μ and unknown variance, with an inverse gamma distribution placed over the variance with parameters $a = \nu/2$ and $b = \nu\sigma^2/2$. In other words, the random variable X is assumed to have a Gaussian distribution with an unknown variance distributed as inverse gamma, and then the variance is marginalized out (integrated out). The reason for the usefulness of this characterization is that the inverse gamma distribution is the conjugate prior distribution of the variance of a Gaussian distribution. As a result, the non-standardized Student's t -distribution arises naturally in many Bayesian inference problems. See below.

Equivalently, this distribution results from compounding a Gaussian distribution with a scaled-inverse-chi-squared distribution with parameters ν and σ^2 . The scaled-inverse-chi-squared distribution is exactly the same distribution as the inverse gamma distribution, but with a different parameterization, i.e. $\nu = a/2$, $\sigma^2 = b/a$.

In terms of inverse scaling parameter λ

An alternative parameterization in terms of an inverse scaling parameter λ (analogous to the way precision is the reciprocal of variance), defined by the relation $\lambda = \sigma^{-2}$. Then the density is defined by

$$p(x|\nu, \mu, \lambda) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x - \mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Other properties of this version of the distribution are:

$$\begin{aligned} E(X) &= \mu && \text{for } \nu > 1, \\ \text{var}(X) &= \frac{1}{\lambda} \frac{\nu}{\nu - 2} && \text{for } \nu > 2, \\ \text{mode}(X) &= \mu. \end{aligned}$$

This distribution results from compounding a Gaussian distribution with mean μ and unknown precision (the reciprocal of the variance), with a gamma distribution placed over the precision with parameters $a = \nu/2$ and $b = \nu/(2\lambda)$. In other words, the random variable X is assumed to have a normal distribution with an unknown precision distributed as gamma, and then this is marginalized over the gamma distribution.

Related distributions

Noncentral t -distribution

The noncentral t -distribution is a different way of generalizing the t -distribution to include a location parameter. Unlike the nonstandardized t -distributions, the noncentral distributions are not symmetric (the median is not the same as the mode).

Discrete Student's t -distribution

The **discrete Student's t -distribution** is defined by its probability mass function at r being proportional to^[8]

$$\prod_{j=1}^k \frac{1}{(r + j + a)^2 + b^2} \quad r = \dots, -1, 0, 1, \dots$$

Here a , b , and k are parameters. This distribution arises from the construction of a system of discrete distributions similar to that of the Pearson distributions for continuous distributions.^[9]

Uses

In frequentist statistical inference

Student's t -distribution arises in a variety of statistical estimation problems where the goal is to estimate an unknown parameter, such as a mean value, in a setting where the data are observed with additive errors. If (as in nearly all practical statistical work) the population standard deviation of these errors is unknown and has to be estimated from the data, the t -distribution is often used to account for the extra uncertainty that results from this estimation. In most such problems, if the standard deviation of the errors were known, a normal distribution would be used instead of the t -distribution.

Confidence intervals and hypothesis tests are two statistical procedures in which the quantiles of the sampling distribution of a particular statistic (e.g. the standard score) are required. In any situation where this statistic is a linear function of the data, divided by the usual estimate of the standard deviation, the resulting quantity can be rescaled and centered to follow Student's t -distribution. Statistical analyses involving means, weighted means, and regression coefficients all lead to statistics having this form.

Quite often, textbook problems will treat the population standard deviation as if it were known and thereby avoid the need to use the Student's t -distribution. These problems are generally of two kinds: (1) those in which the sample size is so large that one may treat a data-based estimate of the variance as if it were certain, and (2) those that illustrate mathematical reasoning, in which the problem of estimating the standard deviation is temporarily ignored because that is not the point that the author or instructor is then explaining.

Hypothesis testing

A number of statistics can be shown to have t -distributions for samples of moderate size under null hypotheses that are of interest, so that the t -distribution forms the basis for significance tests. For example, the distribution of Spearman's rank correlation coefficient ρ , in the null case (zero correlation) is well approximated by the t distribution for sample sizes above about 20 ^[citation needed].

Confidence intervals

Suppose the number A is so chosen that

$$\Pr(-A < T < A) = 0.9,$$

when T has a t -distribution with $n - 1$ degrees of freedom. By symmetry, this is the same as saying that A satisfies

$$\Pr(T < A) = 0.95,$$

so A is the "95th percentile" of this probability distribution, or $A = t_{(0.05, n-1)}$. Then

$$\Pr \left(-A < \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} < A \right) = 0.9,$$

and this is equivalent to

$$\Pr \left(\bar{X}_n - A \frac{S_n}{\sqrt{n}} < \mu < \bar{X}_n + A \frac{S_n}{\sqrt{n}} \right) = 0.9.$$

Therefore the interval whose endpoints are

$$\bar{X}_n \pm A \frac{S_n}{\sqrt{n}}$$

is a 90% confidence interval for μ . Therefore, if we find the mean of a set of observations that we can reasonably expect to have a normal distribution, we can use the t -distribution to examine whether the confidence limits on that mean include some theoretically predicted value – such as the value predicted on a null hypothesis.

It is this result that is used in the Student's t -tests: since the difference between the means of samples from two normal distributions is itself distributed normally, the t -distribution can be used to examine whether that difference can reasonably be supposed to be zero.

If the data are normally distributed, the one-sided $(1 - a)$ -upper confidence limit (UCL) of the mean, can be calculated using the following equation:

$$\text{UCL}_{1-a} = \bar{X}_n + t_{a, n-1} \frac{S_n}{\sqrt{n}}.$$

The resulting UCL will be the greatest average value that will occur for a given confidence interval and population size. In other words, \bar{X}_n being the mean of the set of observations, the probability that the mean of the distribution is inferior to UCL_{1-a} is equal to the confidence level $1 - a$.

Prediction intervals

The t -distribution can be used to construct a prediction interval for an unobserved sample from a normal distribution with unknown mean and variance.

In Bayesian statistics

The Student's t -distribution, especially in its three-parameter (location-scale) version, arises frequently in Bayesian statistics as a result of its connection with the normal distribution. Whenever the variance of a normally distributed random variable is unknown and a conjugate prior placed over it that follows an inverse gamma distribution, the resulting marginal distribution of the variable will follow a Student's t -distribution. Equivalent constructions with the same results involve a conjugate scaled-inverse-chi-squared distribution over the variance, or a conjugate gamma distribution over the precision. If an improper prior proportional to σ^{-2} is placed over the variance, the t -distribution also arises. This is the case regardless of whether the mean of the normally distributed variable is known, is unknown distributed according to a conjugate normally distributed prior, or is unknown distributed according to an improper constant prior.

Related situations that also produce a t -distribution are:

- The marginal posterior distribution of the unknown mean of a normally distributed variable, with unknown prior mean and variance following the above model.
- The prior predictive distribution and posterior predictive distribution of a new normally distributed data point when a series of independent identically distributed normally distributed data points have been observed, with prior mean and variance as in the above model.

Robust parametric modeling

The *t*-distribution is often used as an alternative to the normal distribution as a model for data. It is frequently the case that real data have heavier tails than the normal distribution allows for. The classical approach was to identify outliers and exclude or downweight them in some way. However, it is not always easy to identify outliers (especially in high dimensions), and the *t*-distribution is a natural choice of model for such data and provides a parametric approach to robust statistics.

Lange et al. explored the use of the *t*-distribution for robust modeling of heavy tailed data in a variety of contexts. A Bayesian account can be found in Gelman et al. The degrees of freedom parameter controls the kurtosis of the distribution and is correlated with the scale parameter. The likelihood can have multiple local maxima and, as such, it is often necessary to fix the degrees of freedom at a fairly low value and estimate the other parameters taking this as given. Some authors report that values between 3 and 9 are often good choices. Venables and Ripley suggest that a value of 5 is often a good choice.

Table of selected values

Most statistical textbooks list *t* distribution tables. Nowadays, the better way to a fully precise critical *t* value or a cumulative probability is the statistical function implemented in spreadsheets (Office Excel, OpenOffice Calc, etc.), or an interactive calculating web page. The relevant spreadsheet functions are TDIST and TINV, while online calculating pages save troubles like positions of parameters or names of functions. For example, a MediaWiki page supported by R extension can easily give the interactive result of critical values or cumulative probability, even for noncentral *t*-distribution.

The following table lists a few selected values for *t*-distributions with *v* degrees of freedom for a range of *one-sided* or *two-sided* critical regions. For an example of how to read this table, take the fourth row, which begins with 4; that means *v*, the number of degrees of freedom, is 4 (and if we are dealing, as above, with *n* values with a fixed sum, *n* = 5). Take the fifth entry, in the column headed 95% for *one-sided* (90% for *two-sided*). The value of that entry is "2.132". Then the probability that *T* is less than 2.132 is 95% or $\Pr(-\infty < T < 2.132) = 0.95$; or mean that $\Pr(-2.132 < T < 2.132) = 0.9$.

This can be calculated by the symmetry of the distribution,

$$\Pr(T < -2.132) = 1 - \Pr(T > -2.132) = 1 - 0.95 = 0.05,$$

and so

$$\Pr(-2.132 < T < 2.132) = 1 - 2(0.05) = 0.9.$$

Note that the last row also gives critical points: a *t*-distribution with infinitely many degrees of freedom is a normal distribution. (See Related distributions above).

The first column is the number of degrees of freedom.

<i>One Sided</i>	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
<i>Two Sided</i>	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408

8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

The number at the beginning of each row in the table above is ν which has been defined above as $n - 1$. The percentage along the top is $100\%(1 - \alpha)$. The numbers in the main body of the table are $t_{\alpha, \nu}$. If a quantity T is distributed as a Student's t distribution with ν degrees of freedom, then there is a probability $1 - \alpha$ that T will be less than $t_{\alpha, \nu}$. (Calculated as for a one-tailed or one-sided test, as opposed to a two-tailed test.)

For example, given a sample with a sample variance 2 and sample mean of 10, taken from a sample set of 11 (10 degrees of freedom), using the formula

$$\bar{X}_n \pm A \frac{S_n}{\sqrt{n}}$$

We can determine that at 90% confidence, we have a true mean lying below

$$10 + 1.37218 \frac{\sqrt{2}}{\sqrt{11}} = 10.58510.$$

(In other words, on average, 90% of the times that an upper threshold is calculated by this method, this upper threshold exceeds the true mean.) And, still at 90% confidence, we have a true mean lying over

$$10 - 1.37218 \frac{\sqrt{2}}{\sqrt{11}} = 9.41490.$$

(In other words, on average, 90% of the times that a lower threshold is calculated by this method, this lower threshold lies below the true mean.) So that at 80% confidence (calculated from $1 - 2 \times (1 - 90\%) = 80\%$), we have a true mean lying within the interval

$$\left(10 - 1.37218 \frac{\sqrt{2}}{\sqrt{11}}, 10 + 1.37218 \frac{\sqrt{2}}{\sqrt{11}} \right) = (9.41490, 10.58510).$$

This is generally expressed in interval notation, e.g., for this case, at 80% confidence the true mean is within the interval [9.41490, 10.58510].

(In other words, on average, 80% of the times that upper and lower thresholds are calculated by this method, the true mean is both below the upper threshold and above the lower threshold. This is not the same thing as saying that there is an 80% probability that the true mean lies between a particular pair of upper and lower thresholds that have been calculated by this method—see confidence interval and prosecutor's fallacy.)

For information on the inverse cumulative distribution function see *quantile function*.

Notes

- [1] Hurst, Simon. *The Characteristic Function of the Student-t Distribution* (<http://www.maths.anu.edu.au/research.reports/srr/95/044/>), Financial Mathematics Research Report No. FMRR006-95, Statistics Research Report No. SRR044-95
- [2] "Student" (William Sealy Gosset), original Biometrika paper as a scan (http://www.atmos.washington.edu/~robwood/teaching/451/student_in_biometrika_vol6_no1.pdf)
- [3] Mortimer, Robert G. (2005) *Mathematics for Physical Chemistry*, Academic Press. 3 edition. ISBN 0-12-508347-5 (page 326)
- [4] Walpole, Ronald; Myers, Raymond; Myers, Sharon; Ye, Keying. (2002) *Probability and Statistics for Engineers and Scientists*. Pearson Education, 7th edition, pg. 237 ISBN 81-7758-404-9
- [5] A. Gelman *et al* (1995), *Bayesian Data Analysis*, Chapman & Hall. ISBN 0-412-03991-5. p. 68
- [6] Hogg & Craig (1978, Sections 4.4 and 4.8.)
- [7] See, for example, page 56 of Casella and Berger, *Statistical Inference*, 1990 Duxbury.
- [8] Ord, J.K. (1972) *Families of Frequency Distributions*, Griffin. ISBN 0-85264-137-0 (Table 5.1)
- [9] Ord, J.K. (1972) *Families of Frequency Distributions*, Griffin. ISBN 0-85264-137-0 (Chapter 5)

References

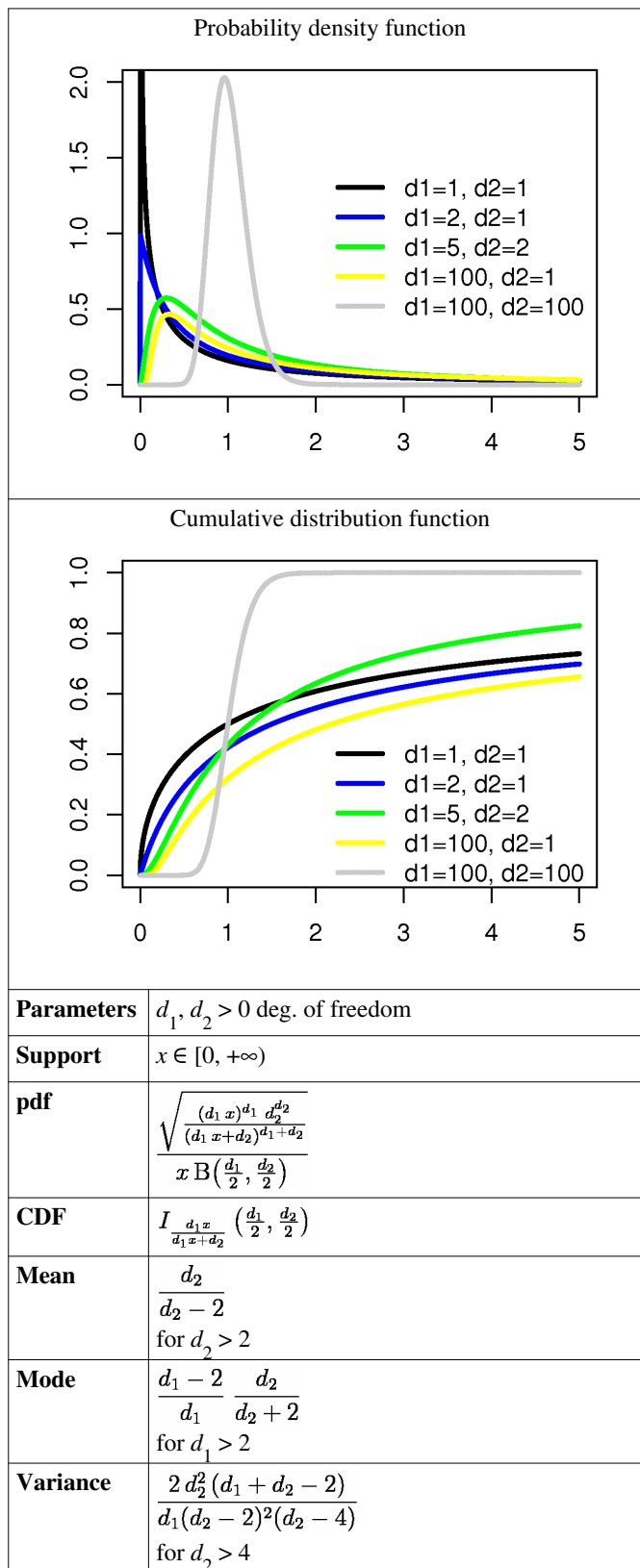
- Senn, S.; Richardson, W. (1994). "The first *t*-test". *Statistics in Medicine* **13** (8): 785–803. doi: 10.1002/sim.4780130802 (<http://dx.doi.org/10.1002/sim.4780130802>). PMID 8047737 (<http://www.ncbi.nlm.nih.gov/pubmed/8047737>).
- Hogg, R.V.; Craig, A.T. (1978). *Introduction to Mathematical Statistics*. New York: Macmillan.
- Venables, W.N.; Ripley, B.D. (2002) *Modern Applied Statistics with S*, Fourth Edition, Springer
- Gelman, Andrew; John B. Carlin, Hal S. Stern, Donald B. Rubin (2003). *Bayesian Data Analysis (Second Edition)* (<http://www.stat.columbia.edu/~gelman/book/>). CRC/Chapman & Hall. ISBN 1-58488-388-X.

External links

- Hazewinkel, Michiel, ed. (2001), "Student distribution" (<http://www.encyclopediaofmath.org/index.php?title=p/s090710>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
 - Earliest Known Uses of Some of the Words of Mathematics (S) (<http://jeff560.tripod.com/s.html>) (*Remarks on the history of the term "Student's distribution"*)
-

F-distribution

Fisher-Snedecor



Skewness	$\frac{(2d_1 + d_2 - 2)\sqrt{8(d_2 - 4)}}{(d_2 - 6)\sqrt{d_1(d_1 + d_2 - 2)}}$ for $d_2 > 6$
Ex. kurtosis	<i>see text</i>
MGF	<i>does not exist, raw moments defined in text and in</i>
CF	<i>see text</i>

In probability theory and statistics, the **F-distribution** is a continuous probability distribution.^[1] It is also known as **Snedecor's F distribution** or the **Fisher-Snedecor distribution** (after R.A. Fisher and George W. Snedecor). The F-distribution arises frequently as the null distribution of a test statistic, most notably in the analysis of variance; see F-test.

Definition

If a random variable X has an F-distribution with parameters d_1 and d_2 , we write $X \sim F(d_1, d_2)$. Then the probability density function for X is given by

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

$$= \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2} - 1} \left(1 + \frac{d_1}{d_2} x\right)^{-\frac{d_1 + d_2}{2}}$$

for real $x \geq 0$. Here B is the beta function. In many applications, the parameters d_1 and d_2 are positive integers, but the distribution is well-defined for positive real values of these parameters.

The cumulative distribution function is

$$F(x; d_1, d_2) = I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right),$$

where I is the regularized incomplete beta function.

The expectation, variance, and other details about the $F(d_1, d_2)$ are given in the sidebox; for $d_2 > 8$, the excess kurtosis is

$$\gamma_2 = 12 \frac{d_1(5d_2 - 22)(d_1 + d_2 - 2) + (d_2 - 4)(d_2 - 2)^2}{d_1(d_2 - 6)(d_2 - 8)(d_1 + d_2 - 2)}.$$

The k -th moment of an $F(d_1, d_2)$ distribution exists and is finite only when $2k < d_2$ and it is equal to

$$\mu_X(k) = \left(\frac{d_2}{d_1}\right)^k \frac{\Gamma\left(\frac{d_1}{2} + k\right) \Gamma\left(\frac{d_2}{2} - k\right)}{\Gamma\left(\frac{d_1}{2}\right) \Gamma\left(\frac{d_2}{2}\right)}$$

The F -distribution is a particular parametrization of the beta prime distribution, which is also called the beta distribution of the second kind.

The characteristic function is listed incorrectly in many standard references (e.g.,). The correct expression^[2] is

$$\varphi_{d_1, d_2}^F(s) = \frac{\Gamma\left(\frac{d_1 + d_2}{2}\right)}{\Gamma\left(\frac{d_2}{2}\right)} U\left(\frac{d_1}{2}, 1 - \frac{d_2}{2}, -\frac{d_2}{d_1}is\right)$$

where $U(a, b, z)$ is the confluent hypergeometric function of the second kind.

Characterization

A random variate of the F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled chi-squared variates:^[3]

$$X = \frac{U_1/d_1}{U_2/d_2}$$

where

- U_1 and U_2 have chi-squared distributions with d_1 and d_2 degrees of freedom respectively, and
- U_1 and U_2 are independent.

In instances where the F-distribution is used, for example in the analysis of variance, independence of U_1 and U_2 might be demonstrated by applying Cochran's theorem.

Equivalently, the random variable of the F-distribution may also be written

$$X = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

where s_1^2 and s_2^2 are the sums of squares S_1^2 and S_2^2 from two normal processes with variances σ_1^2 and σ_2^2 divided by the corresponding number of χ^2 degrees of freedom, d_1 and d_2 respectively.

In a Frequentist context, a scaled F-distribution therefore gives the probability $p(s_1^2/s_2^2 | \sigma_1^2, \sigma_2^2)$, with the F-distribution itself, without any scaling, applying where σ_1^2 is being taken equal to σ_2^2 . This is the context in which the F-distribution most generally appears in F-tests: where the null hypothesis is that two independent normal variances are equal, and the observed sums of some appropriately selected squares are then examined to see whether their ratio is significantly incompatible with this null hypothesis.

The quantity X has the same distribution in Bayesian statistics, if an uninformative rescaling-invariant Jeffreys prior is taken for the prior probabilities of σ_1^2 and σ_2^2 .^[4] In this context, a scaled F-distribution thus gives the posterior probability $p(\sigma_2^2/\sigma_1^2 | s_1^2, s_2^2)$, where now the observed sums s_1^2 and s_2^2 are what are taken as known.

Generalization

A generalization of the (central) F-distribution is the noncentral F-distribution.

Related distributions and properties

- If $X \sim \chi_{d_1}^2$ and $Y \sim \chi_{d_2}^2$ are independent, then $\frac{X/d_1}{Y/d_2} \sim F(d_1, d_2)$
- If $X \sim \text{Beta}(d_1/2, d_2/2)$ (Beta distribution) then $\frac{d_2 X}{d_1(1-X)} \sim F(d_1, d_2)$
- Equivalently, if $X \sim F(d_1, d_2)$, then $\frac{d_1 X/d_2}{1 + d_1 X/d_2} \sim \text{Beta}(d_1/2, d_2/2)$.
- If $X \sim F(d_1, d_2)$ then $Y = \lim_{d_2 \rightarrow \infty} d_1 X$ has the chi-squared distribution $\chi_{d_1}^2$
- $F(d_1, d_2)$ is equivalent to the scaled Hotelling's T-squared distribution $\frac{d_2}{d_1(d_1 + d_2 - 1)} T^2(d_1, d_1 + d_2 - 1)$
- If $X \sim F(d_1, d_2)$ then $X^{-1} \sim F(d_2, d_1)$.
- If $X \sim t(n)$ then

$$X^2 \sim F(1, n)$$

$$X^{-2} \sim F(n, 1)$$

- F-distribution is a special case of type 6 Pearson distribution

- If X and Y are independent, with $X, Y \sim \text{Laplace}(\mu, b)$ then

$$\frac{|X-\mu|}{|Y-\mu|} \sim F(2, 2)$$

- If $X \sim F(n, m)$ then $\frac{\log X}{2} \sim \text{FisherZ}(n, m)$ (Fisher's z-distribution)
- The noncentral F-distribution simplifies to the F-distribution if $\lambda = 0$.
- The doubly noncentral F-distribution simplifies to the F-distribution if $\lambda_1 = \lambda_2 = 0$
- If $Q_X(p)$ is the quantile p for $X \sim F(d_1, d_2)$ and $Q_Y(1-p)$ is the quantile $1-p$ for $Y \sim F(d_2, d_1)$, then

$$Q_X(p) = \frac{1}{Q_Y(1-p)}.$$

References

- [1] NIST (2006). Engineering Statistics Handbook - F Distribution (<http://www.itl.nist.gov/div898/handbook/eda/section3/eda3665.htm>)
- [2] Phillips, P. C. B. (1982) "The true characteristic function of the F distribution," *Biometrika*, 69: 261-264
- [3] M.H. DeGroot (1986), *Probability and Statistics* (2nd Ed), Addison-Wesley. ISBN 0-201-11366-X, p. 500
- [4] G.E.P. Box and G.C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley. p.110

External links

- Table of critical values of the F-distribution (<http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm>)
- Earliest Uses of Some of the Words of Mathematics: entry on F-distribution contains a brief history (<http://jeff560.tripod.com/f.html>)
- Free calculator for F-testing (<http://www.waterlog.info/f-test.htm>)

Feature scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Motivation

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Methods

Rescaling

The simplest method is rescaling the range of features to make the features independent of each other and aims to scale the range in $[0, 1]$ or $[-1, 1]$. Selecting the target range depends on the nature of the data. The general formula is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x is an original value, x' is the normalized value. For example, suppose that we have the students' weight data, and the students' weights span [160 pounds, 200 pounds]. To rescale this data, we first subtract 160 from each

student's weight and divide the result by 40 (the difference between the maximum and minimum weights).

Standardization

In machine learning, we can handle various types of data, e.g., audio signals, pixel values for image data, and etc., and this data can include multiple dimensions. Feature standardization makes the values of each feature in the data have zero-mean and unit-variance. This method is widely used for normalization in many machine learning algorithms (e.g., support vector machines, logistic regression, and neural networks). In general, we first calculate the mean and standard deviation for each feature, and then, subtract the mean in each feature. Then, we divide the values (mean is already subtracted) of each feature by its standard deviation.

Scaling to unit length

Another option that is widely used in machine-learning is to scale the components of a feature vector such that the complete vector has length one. This usually means dividing each component by the Euclidean length of the vector. In some applications (e.g. Histogram features) it can be more practical to use the L1 norm (i.e. Manhattan or City-Block Length) of the feature vector:

$$x' = \frac{x}{\|x\|}$$

This is especially important if in the following learning steps the Scalar Metric is used as a distance measure.

Application

In gradient descent, feature scaling can improve the convergence speed of the algorithm. In SVM, it reduces the time to find support vectors and helps the data points be properly placed in the space of kernel function.

References

- S. Aksoy and R. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognit. Lett.*, Special Issue on Image and Video Retrieval, 2000 http://www.cs.bilkent.edu.tr/~saksoy/papers/prletters01_likelihood.pdf
- S. Tsakalidis, V. Doumptotis & W. Byrne, "Discriminative Linear Transforms for Feature Normalization and Speaker Adaptation in HMM Estimation", *Proc. ICSLP'02, Denver*. http://malach.umiacs.umd.edu/pubs/VD_05_Discrim_linear.pdf
- Liefeng Bo, Ling Wang, and Licheng Jiao, "Feature Scaling for Kernel Fisher Discriminant Analysis Using Leave-one-out Cross Validation", *Neural Computation (NECO)*, vol. 18(4), pp. 961–978, 2006 <http://www.cs.washington.edu/homes/lfb/paper/nc06.pdf>
- A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification," in *Proc. ICASSP, Las Vegas, Apr. 2008*. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4517925
- E. Youn, M. K. Jeong, "Class dependent feature scaling method using naive Bayes classifier for text datamining" *Pattern Recognition Letters*, 2009. <http://www.sciencedirect.com/science/article/pii/S0167865508003553>
- S. Theodoridis, K. Koutroubas. (2008) "Pattern Recognition", Academic Press, 4 edition, ISBN 978-1-59749-272-0

External links

- Lecture by Andrew Ng on feature scaling (<http://openclassroom.stanford.edu/MainFolder/VideoPage.php?course=MachineLearning&video=03.1-LinearRegressionII-FeatureScaling&speed=100/>)
 - Gradient Descent using feature scaling (<http://www.statalgo.com/2011/10/17/stanford-ml-1-2-gradient-descent/>)
 - Feature normalization ([http://mipa.med.upatras.gr/educational resources/Data Normalization.pdf](http://mipa.med.upatras.gr/educational%20resources/Data%20Normalization.pdf))
-

Correlation and Regression

Covariance

In probability theory and statistics, **covariance** is a measure of how much two random variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e., the variables tend to show similar behavior, the covariance is positive.^[1] In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e., the variables tend to show opposite behavior, the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables. The magnitude of the covariance is not easy to interpret. The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

A distinction must be made between (1) the covariance of two random variables, which is a population parameter that can be seen as a property of the joint probability distribution, and (2) the sample covariance, which serves as an estimated value of the parameter.

Definition

The covariance between two jointly distributed real-valued random variables x and y with finite second moments is defined^[2] as

$$\sigma(x, y) = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])],$$

where $\mathbb{E}[x]$ is the expected value of x , also known as the mean of x . By using the linearity property of expectations, this can be simplified to

$$\begin{aligned} \sigma(x, y) &= \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \\ &= \mathbb{E}[xy - x\mathbb{E}[y] - \mathbb{E}[x]y + \mathbb{E}[x]\mathbb{E}[y]] \\ &= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] + \mathbb{E}[x]\mathbb{E}[y] \\ &= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]. \end{aligned}$$

However, when $\mathbb{E}[xy] \approx \mathbb{E}[x]\mathbb{E}[y]$, this last equation is prone to catastrophic cancellation when computed with floating point arithmetic and thus should be avoided in computer programs when the data has not been centered before.^[3]

For random vectors \mathbf{X} and \mathbf{Y} (of dimension m and n respectively) the $m \times n$ cross covariance matrix (also known as **dispersion matrix** or **variance–covariance matrix**,^[4] or simply called covariance matrix) is equal to

$$\begin{aligned} \sigma(\mathbf{x}, \mathbf{y}) &= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^T] \\ &= \mathbb{E}[\mathbf{xy}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}]^T, \end{aligned}$$

where \mathbf{m}^T is the transpose of the vector (or matrix) \mathbf{m} .

The (i, j) -th element of this matrix is equal to the covariance $\text{Cov}(x_i, y_j)$ between the i -th scalar component of x and the j -th scalar component of y . In particular, $\text{Cov}(y, x)$ is the transpose of $\text{Cov}(x, y)$.

For a vector $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_m]^T$ of m jointly distributed random variables with finite second moments, its covariance matrix is defined as

$$\Sigma(\mathbf{x}) = \sigma(\mathbf{x}, \mathbf{x}).$$

Random variables whose covariance is zero are called uncorrelated.

The units of measurement of the covariance $\text{Cov}(x, y)$ are those of x times those of y . By contrast, correlation coefficients, which depend on the covariance, are a dimensionless measure of linear dependence. (In fact, correlation coefficients can simply be understood as a normalized version of covariance.)

Properties

- Variance is a special case of the covariance when the two variables are identical:

$$\sigma(x, x) = \sigma^2(x).$$

- If $x, y, w,$ and v are real-valued random variables and a, b, c, d are constant ("constant" in this context means non-random), then the following facts are a consequence of the definition of covariance:

$$\sigma(x, a) = 0$$

$$\sigma(x, x) = \sigma^2(x)$$

$$\sigma(x, y) = \sigma(y, x)$$

$$\sigma(ax, by) = ab \sigma(x, y)$$

$$\sigma(x + a, y + b) = \sigma(x, y)$$

$$\sigma(ax + by, cw + dv) = ac \sigma(x, w) + ad \sigma(x, v) + bc \sigma(y, w) + bd \sigma(y, v)$$

For a sequence x_1, \dots, x_n of random variables, and constants a_1, \dots, a_n , we have

$$\sigma^2\left(\sum_{i=1}^n a_i x_i\right) = \sum_{i=1}^n a_i^2 \sigma^2(x_i) + 2 \sum_{i,j:i < j} a_i a_j \sigma(x_i, x_j) = \sum_{i,j} a_i a_j \sigma(x_i, x_j)$$

A more general identity for covariance matrices

Let \mathbf{x} be a random vector with covariance matrix $\Sigma(\mathbf{x})$, and let A be a matrix that can act on \mathbf{x} . The covariance matrix of the vector $A\mathbf{x}$ is:

$$\Sigma(A\mathbf{x}) = A \Sigma(\mathbf{x}) A^T.$$

This is a direct result of the linearity of expectation and is useful when applying a linear transformation, such as a whitening transformation, to a vector.

Uncorrelatedness and independence

If x and y are independent, then their covariance is zero. This follows because under independence,

$$E[xy] = E[x] \cdot E[y].$$

The converse, however, is not generally true. For example, let x be uniformly distributed in $[-1, 1]$ and let $y = x^2$. Clearly, x and y are dependent, but

$$\begin{aligned} \sigma(x, y) &= \sigma(x, x^2) \\ &= E[x \cdot x^2] - E[x] \cdot E[x^2] \\ &= E[x^3] - E[x]E[x^2] \\ &= 0 - 0 \cdot E[x^2] \\ &= 0. \end{aligned}$$

In this case, the relationship between y and x is non-linear, while correlation and covariance are measures of linear dependence between two variables. This example shows that if two variables are uncorrelated, that does not in general imply that they are independent. However, if two variables are jointly normally distributed (but not if they are merely individually normally distributed), uncorrelatedness *does* imply independence.

Relationship to inner products

Many of the properties of covariance can be extracted elegantly by observing that it satisfies similar properties to those of an inner product:

1. bilinear: for constants a and b and random variables x, y, z , $\sigma(ax + by, z) = a \sigma(x, z) + b \sigma(y, z)$;
2. symmetric: $\sigma(x, y) = \sigma(y, x)$;
3. positive semi-definite: $\sigma^2(x) = \sigma(x, x) \geq 0$ for all random variables x , and $\sigma(x, x) = 0$ implies that x is a constant random variable (K).

In fact these properties imply that the covariance defines an inner product over the quotient vector space obtained by taking the subspace of random variables with finite second moment and identifying any two that differ by a constant. (This identification turns the positive semi-definiteness above into positive definiteness.) That quotient vector space is isomorphic to the subspace of random variables with finite second moment and mean zero; on that subspace, the covariance is exactly the L^2 inner product of real-valued functions on the sample space.

As a result for random variables with finite variance, the inequality

$$|\sigma(x, y)| \leq \sigma(x)\sigma(y)$$

holds via the Cauchy–Schwarz inequality.

Proof: If $\sigma^2(y) = 0$, then it holds trivially. Otherwise, let random variable

$$z = x - \frac{\sigma(x, y)}{\sigma^2(y)}y.$$

Then we have

$$\begin{aligned} 0 \leq \sigma^2(z) &= \sigma\left(x - \frac{\sigma(x, y)}{\sigma^2(y)}y, x - \frac{\sigma(x, y)}{\sigma^2(y)}y\right) \\ &= \sigma^2(x) - \frac{(\sigma(x, y))^2}{\sigma^2(y)}. \end{aligned}$$

Calculating the sample covariance

The sample covariance of N observations of K variables is the K -by- K matrix $\bar{q} = [[q_{jk}]]$ with the entries

$$q_{jk} = \frac{1}{N-1} \sum_{i=1}^N (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k),$$

which is an estimate of the covariance between variable j and variable k .

The sample mean and the sample covariance matrix are unbiased estimates of the mean and the covariance matrix of the random vector \mathbf{x} , a row vector whose j^{th} element ($j = 1, \dots, K$) is one of the random variables. The reason the sample covariance matrix has $N - 1$ in the denominator rather than N is essentially that the population mean $E(\mathbf{x})$ is not known and is replaced by the sample mean $\bar{\mathbf{x}}$. If the population mean $E(\mathbf{x})$ is known, the analogous unbiased estimate is given by

$$q_{jk} = \frac{1}{N} \sum_{i=1}^N (x_{ij} - E(x_j))(x_{ik} - E(x_k))$$

Comments

The covariance is sometimes called a measure of "linear dependence" between the two random variables. That does not mean the same thing as in the context of linear algebra (see linear dependence). When the covariance is normalized, one obtains the correlation coefficient. From it, one can obtain the Pearson coefficient, which gives us the goodness of the fit for the best possible linear function describing the relation between the variables. In this sense covariance is a linear gauge of dependence.

Applications

In genetics and molecular biology

Covariance is an important measure in biology. Certain sequences of DNA are conserved more than others among species, and thus to study secondary and tertiary structures of proteins, or of RNA structures, we compare sequences in closely related species. If we find sequence changes or no changes at all in noncoding RNA (such as microRNA), we can find out about which sequences are necessary for common structural motifs, such as an RNA loop.

In financial economics

Covariances play a key role in financial economics, especially in portfolio theory and in the capital asset pricing model. Covariances among various assets' returns are used to determine, under certain assumptions, the relative amounts of different assets that investors should (in a normative analysis) or are predicted to (in a positive analysis) choose to hold in a context of diversification.

References

- [1] <http://mathworld.wolfram.com/Covariance.html>
- [2] Oxford Dictionary of Statistics, Oxford University Press, 2002, p. 104.
- [3] Donald E. Knuth (1998). *The Art of Computer Programming*, volume 2: *Seminumerical Algorithms*, 3rd edn., p. 232. Boston: Addison-Wesley.
- [4] W. J. Krzanowski, *Principles of Multivariate Analysis*, Chap. 7.1, Oxford University Press, New York, 1988

External links

- Hazewinkel, Michiel, ed. (2001), "Covariance" (<http://www.encyclopediaofmath.org/index.php?title=p/c026800>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- MathWorld page on calculating the sample covariance (<http://mathworld.wolfram.com/Covariance.html>)
- Covariance Tutorial using R (<http://www.r-tutor.com/elementary-statistics/numerical-measures/covariance>)

Correlation and dependence

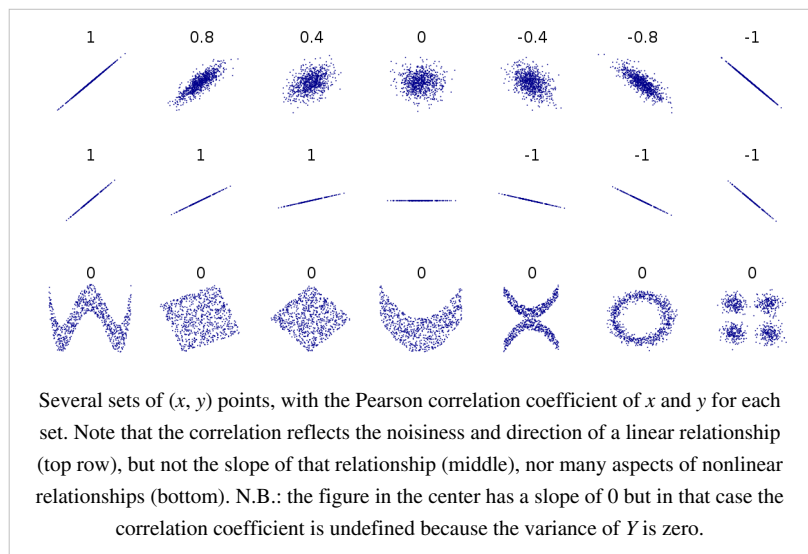
In statistics, **dependence** is any statistical relationship between two random variables or two sets of data. **Correlation** refers to any of a broad class of statistical relationships involving dependence.

Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling; however, statistical dependence is not sufficient to demonstrate the presence of such a causal relationship (i.e., correlation does not imply causation).

Formally, *dependence* refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. In loose usage, *correlation* can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values. There are several **correlation coefficients**, often denoted ρ or r , measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function of the other). Other correlation coefficients have been developed to be more robust than the Pearson correlation – that is, more sensitive to nonlinear relationships.^{[1][2][3]} Mutual information can also be applied to measure dependence between two variables.

Pearson's product-moment coefficient

The most familiar measure of dependence between two quantities is the Pearson product-moment correlation coefficient, or "Pearson's correlation coefficient", commonly called simply "the correlation coefficient". It is obtained by dividing the covariance of the two variables by the product of their standard deviations. Karl Pearson developed the coefficient from a similar but slightly different idea by Francis Galton.^[4]



The population correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

where E is the expected value operator, *cov* means covariance, and, *corr* a widely used alternative notation for the correlation coefficient.

The Pearson correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the Cauchy–Schwarz inequality that the correlation cannot exceed 1 in absolute value. The correlation coefficient is symmetric: $\text{corr}(X, Y) = \text{corr}(Y, X)$.

The Pearson correlation is +1 in the case of a perfect direct(increasing) linear relationship (correlation), -1 in the case of a perfect decreasing (inverse) linear relationship (**anticorrelation**),^[5] and some value between -1 and 1 in all other cases, indicating the degree of linear dependence between the variables. As it approaches zero there is less of a relationship (closer to uncorrelated). The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables.

If the variables are independent, Pearson's correlation coefficient is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables. For example, suppose the random variable X is symmetrically distributed about zero, and $Y = X^2$. Then Y is completely determined by X , so that X and Y are perfectly dependent, but their correlation is zero; they are uncorrelated. However, in the special case when X and Y are jointly normal, uncorrelatedness is equivalent to independence.

If we have a series of n measurements of X and Y written as x_i and y_i where $i = 1, 2, \dots, n$, then the *sample correlation coefficient* can be used to estimate the population Pearson correlation r between X and Y . The sample correlation coefficient is written

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

where \bar{x} and \bar{y} are the sample means of X and Y , and s_x and s_y are the sample standard deviations of X and Y .

This can also be written as:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

If x and y are results of measurements that contain measurement error, the realistic limits on the correlation coefficient are not -1 to +1 but a smaller range.

For the case of a linear model with a single independent variable, the coefficient of determination (R squared) is the square of r , Pearson's product-moment coefficient .

Rank correlation coefficients

Rank correlation coefficients, such as Spearman's rank correlation coefficient and Kendall's rank correlation coefficient (τ) measure the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship. If, as the one variable increases, the other *decreases*, the rank correlation coefficients will be negative. It is common to regard these rank correlation coefficients as alternatives to Pearson's coefficient, used either to reduce the amount of calculation or to make the coefficient less sensitive to non-normality in distributions. However, this view has little mathematical basis, as rank correlation coefficients measure a different type of relationship than the Pearson product-moment correlation coefficient, and are best seen as measures of a different type of association, rather than as alternative measure of the population correlation coefficient.^{[6][7]}

To illustrate the nature of rank correlation, and its difference from linear correlation, consider the following four pairs of numbers (x, y) :

$$(0, 1), (10, 100), (101, 500), (102, 2000).$$

As we go from each pair to the next pair x increases, and so does y . This relationship is perfect, in the sense that an increase in x is *always* accompanied by an increase in y . This means that we have a perfect rank correlation, and both Spearman's and Kendall's correlation coefficients are 1, whereas in this example Pearson product-moment correlation coefficient is 0.7544, indicating that the points are far from lying on a straight line. In the same way if y always *decreases* when x *increases*, the rank correlation coefficients will be -1, while the Pearson product-moment

correlation coefficient may or may not be close to -1 , depending on how close the points are to a straight line. Although in the extreme cases of perfect rank correlation the two coefficients are both equal (being both $+1$ or both -1) this is not in general so, and values of the two coefficients cannot meaningfully be compared. For example, for the three pairs (1, 1) (2, 3) (3, 2) Spearman's coefficient is $1/2$, while Kendall's coefficient is $1/3$.

Other measures of dependence among random variables

The information given by a correlation coefficient is not enough to define the dependence structure between random variables. The correlation coefficient completely defines the dependence structure only in very particular cases, for example when the distribution is a multivariate normal distribution. (See diagram above.) In the case of elliptical distributions it characterizes the (hyper-)ellipses of equal density, however, it does not completely characterize the dependence structure (for example, a multivariate t-distribution's degrees of freedom determine the level of tail dependence).

Distance correlation and Brownian covariance / Brownian correlation ^{[8][9]} were introduced to address the deficiency of Pearson's correlation that it can be zero for dependent random variables; zero distance correlation and zero Brownian correlation imply independence.

The correlation ratio is able to detect almost any functional dependency ^[citation needed] Wikipedia:Please clarify, and the entropy-based mutual information, total correlation and dual total correlation are capable of detecting even more general dependencies. These are sometimes referred to as multi-moment correlation measures ^[citation needed], in comparison to those that consider only second moment (pairwise or quadratic) dependence.

The polychoric correlation is another correlation applied to ordinal data that aims to estimate the correlation between theorised latent variables.

One way to capture a more complete view of dependence structure is to consider a copula between them.

The coefficient of determination generalizes the correlation coefficient for relationships beyond simple linear regression.

Sensitivity to the data distribution

The degree of dependence between variables X and Y does not depend on the scale on which the variables are expressed. That is, if we are analyzing the relationship between X and Y , most correlation measures are unaffected by transforming X to $a + bX$ and Y to $c + dY$, where a , b , c , and d are constants. This is true of some correlation statistics as well as their population analogues. Some correlation statistics, such as the rank correlation coefficient, are also invariant to monotone transformations of the marginal distributions of X and/or Y .

Most correlation measures are sensitive to the manner in which X and Y are sampled. Dependencies tend to be stronger if viewed over a wider range of values. Thus, if we consider the correlation coefficient between the heights of fathers and their sons over all adult males, and compare it to the same correlation coefficient calculated when the fathers are selected to be between 165 cm and 170 cm in height, the correlation will be weaker in the latter case. Several techniques have been developed that attempt to correct for range restriction in one or both variables, and are commonly used in meta-analysis; the most common are Thorndike's case II and case III equations.

Various correlation measures in use may be undefined for certain joint distributions of X and Y .

For example, the Pearson correlation coefficient is defined in terms of moments, and hence will be undefined if the moments are undefined. Measures of dependence based on quantiles are always defined. Sample-based statistics intended to estimate population measures of dependence may or may not have desirable statistical properties such as being unbiased, or asymptotically consistent, based on the spatial structure of the population from which the data were sampled.

Sensitivity to the data distribution can be used to an advantage. For example, scaled correlation is designed to use the sensitivity to the range in order to pick out correlations between fast components of time series.^[10] By reducing the range of values in a controlled manner, the correlations on long time scale are filtered out and only the correlations on short time scales are revealed.

Correlation matrices

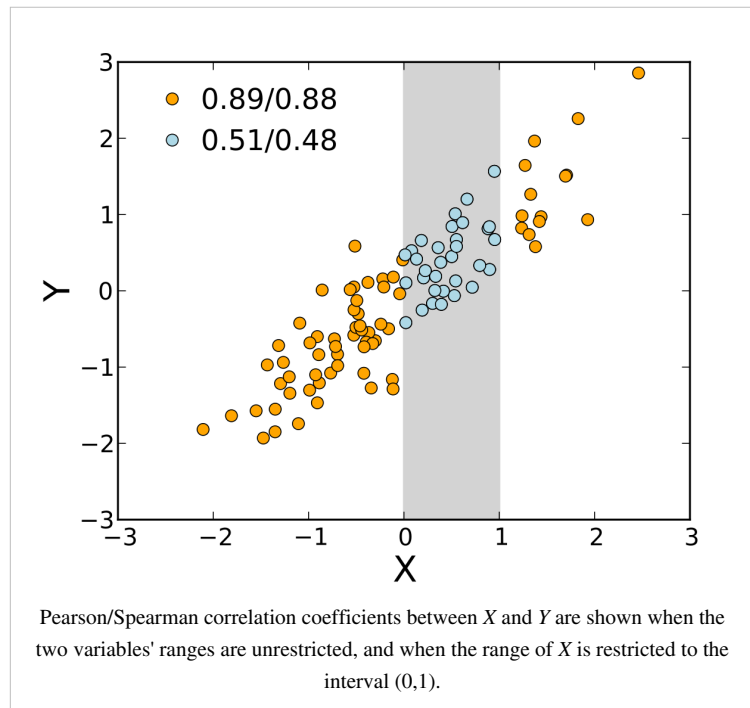
The correlation matrix of n random variables X_1, \dots, X_n is the $n \times n$ matrix whose i,j entry is $\text{corr}(X_i, X_j)$. If the measures of correlation used are product-moment coefficients, the correlation matrix is the same as the covariance matrix of the standardized random variables $X_i / \sigma(X_i)$ for $i = 1, \dots, n$. This applies to both the matrix of population correlations (in which case " σ " is the population standard deviation), and to the matrix of sample correlations (in which case " σ " denotes the sample standard deviation). Consequently, each is necessarily a positive-semidefinite matrix.

The correlation matrix is symmetric because the correlation between X_i and X_j is the same as the correlation between X_j and X_i .

Common misconceptions

Correlation and causality

The conventional dictum that "correlation does not imply causation" means that correlation cannot be used to infer a causal relationship between the variables. This dictum should not be taken to mean that correlations cannot indicate the potential existence of causal relations. However, the causes underlying the correlation, if any, may be indirect and unknown, and high correlations also overlap with identity relations (tautologies), where no causal process exists.



Consequently, establishing a correlation between two variables is not a sufficient condition to establish a causal relationship (in either direction).

A correlation between age and height in children is fairly causally transparent, but a correlation between mood and health in people is less so. Does improved mood lead to improved health, or does good health lead to good mood, or both? Or does some other factor underlie both? In other words, a correlation can be taken as evidence for a possible causal relationship, but cannot indicate what the causal relationship, if any, might be.

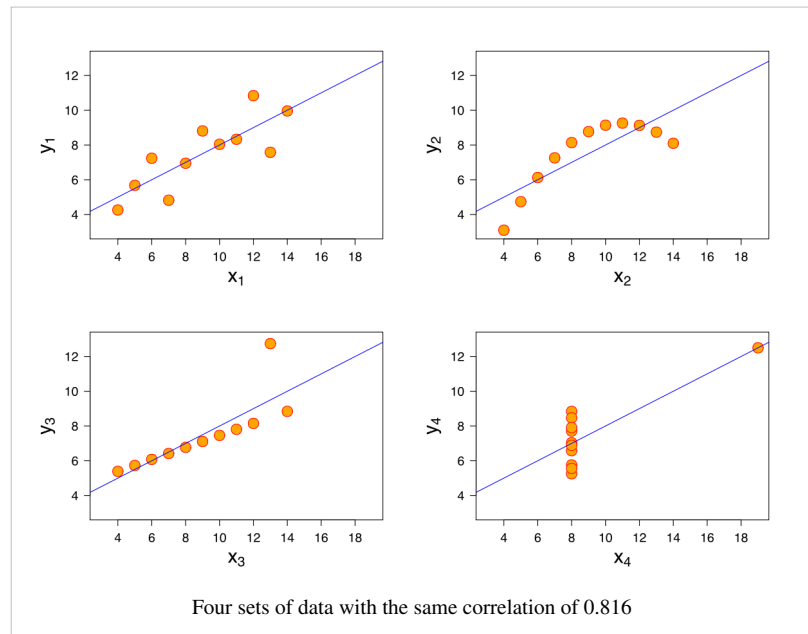
Correlation and linearity

The Pearson correlation coefficient indicates the strength of a linear relationship between two variables, but its value generally does not completely characterize their relationship. In particular, if the conditional mean of Y given X , denoted $E(Y|X)$, is not linear in X , the correlation coefficient will not fully determine the form of $E(Y|X)$.

The image on the right shows scatterplots of Anscombe's quartet, a set of four different pairs of variables created by Francis Anscombe. The four y variables have the same mean (7.5), variance (4.12), correlation (0.816) and regression line

($y = 3 + 0.5x$). However, as can be seen on the plots, the distribution of the variables is very different. The first one (top left) seems to be distributed normally, and corresponds to what one would expect when considering two variables correlated and following the assumption of normality. The second one (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear. In this case the Pearson correlation coefficient does not indicate that there is an exact functional relationship: only the extent to which that relationship can be approximated by a linear relationship. In the third case (bottom left), the linear relationship is perfect, except for one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816. Finally, the fourth example (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

These examples indicate that the correlation coefficient, as a summary statistic, cannot replace visual examination of the data. Note that the examples are sometimes said to demonstrate that the Pearson correlation assumes that the data follow a normal distribution, but this is not correct.



Life-time of correlation

Most analyses do not take into account variation of the correlation coefficient with time. If the variables are non-stationary, then some concepts of choosing optimal time intervals are needed. The durability of correlation should also be calculated in such a case.^[11]

Bivariate normal distribution

If a pair (X, Y) of random variables follows a bivariate normal distribution, the conditional mean $E(X|Y)$ is a linear function of Y , and the conditional mean $E(Y|X)$ is a linear function of X . The correlation coefficient r between X and Y , along with the marginal means and variances of X and Y , determines this linear relationship:

$$E(Y | X) = E(Y) + r\sigma_y \frac{X - E(X)}{\sigma_x},$$

where $E(X)$ and $E(Y)$ are the expected values of X and Y , respectively, and σ_x and σ_y are the standard deviations of X and Y , respectively.

Partial correlation

If a population or data-set is characterized by more than two variables, a partial correlation coefficient measures the strength of dependence between a pair of variables that is not accounted for by the way in which they both change in response to variations in a selected subset of the other variables.

References

- [1] Croxton, Frederick Emory; Cowden, Dudley Johnstone; Klein, Sidney (1968) *Applied General Statistics*, Pitman. ISBN 9780273403159 (page 625)
- [2] Dietrich, Cornelius Frank (1991) *Uncertainty, Calibration and Probability: The Statistics of Scientific and Industrial Measurement* 2nd Edition, A. Higler. ISBN 9780750300605 (Page 331)
- [3] Aitken, Alexander Craig (1957) *Statistical Mathematics* 8th Edition. Oliver & Boyd. ISBN 9780050013007 (Page 95)
- [4] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient (<http://www.jstor.org/stable/2685263>). *The American Statistician*, 42(1):59–66, February 1988.
- [5] Dowdy, S. and Wearden, S. (1983). "Statistics for Research", Wiley. ISBN 0-471-08602-9 pp 230
- [6] Yule, G.U and Kendall, M.G. (1950), "An Introduction to the Theory of Statistics", 14th Edition (5th Impression 1968). Charles Griffin & Co. pp 258–270
- [7] Kendall, M. G. (1955) "Rank Correlation Methods", Charles Griffin & Co.
- [8] Székely, G. J. Rizzo, M. L. and Bakirov, N. K. (2007). "Measuring and testing independence by correlation of distances", *Annals of Statistics*, 35/6, 2769–2794. Reprint (<http://personal.bgsu.edu/~mrizzo/energy/AOS0283-reprint.pdf>)
- [9] Székely, G. J. and Rizzo, M. L. (2009). "Brownian distance covariance", *Annals of Applied Statistics*, 3/4, 1233–1303. Reprint (<http://personal.bgsu.edu/~mrizzo/energy/AOAS312.pdf>)
- [10] Nikolić D, Muresan RC, Feng W, Singer W (2012) Scaled correlation analysis: a better way to compute a cross-correlogram. *European Journal of Neuroscience*, pp. 1–21,
- [11] Buda, Andrzej; Jarynowski, Andrzej (2010) *Life-time of correlations and its applications*, [Wrocław] : Wydawnictwo Niezależne http://th.if.uj.edu.pl/~gulakov/life_corr/

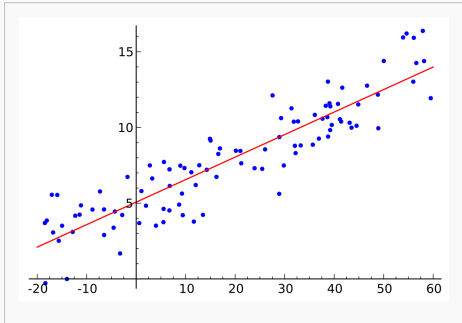
Further reading


- Cohen, J., Cohen P., West, S.G., & Aiken, L.S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.)*. Psychology Press. ISBN 0-8058-2223-2.
- Hazewinkel, Michiel, ed. (2001), "Correlation (in statistics)" (<http://www.encyclopediaofmath.org/index.php?title=p/c026560>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4

External links

- MathWorld page on (cross-) correlation coefficient(s) of a sample. (<http://mathworld.wolfram.com/CorrelationCoefficient.html>)
- Compute Significance between two correlations (<http://peaks.informatik.uni-erlangen.de/cgi-bin/usignificance.cgi>) – A useful website if one wants to compare two correlation values.
- A MATLAB Toolbox for computing Weighted Correlation Coefficients (<http://www.mathworks.com/matlabcentral/fileexchange/20846>)
- Proof that the Sample Bivariate Correlation Coefficient has Limits ± 1 ([http://www.docstoc.com/docs/3530180/Proof-that-the-Sample-Bivariate-Correlation-Coefficient-has-Limits-\(Plus-or-Minus\)-1](http://www.docstoc.com/docs/3530180/Proof-that-the-Sample-Bivariate-Correlation-Coefficient-has-Limits-(Plus-or-Minus)-1))
- Interactive Flash simulation on the correlation of two normally distributed variables. (http://nagysandor.eu/AsimovTeka/correlation_en/index.html) Author: Juha Puranen.

Regression analysis

Regression analysis	
	
Models	
<ul style="list-style-type: none"> • Linear regression • Simple regression • Ordinary least squares • Polynomial regression • General linear model 	
<ul style="list-style-type: none"> • Generalized linear model • Discrete choice • Logistic regression • Multinomial logit • Mixed logit • Probit • Multinomial probit • Ordered logit • Ordered probit 	

• Poisson
• Multilevel model
• Fixed effects
• Random effects
• Mixed model
• Nonlinear regression
• Nonparametric
• Semiparametric
• Robust
• Quantile
• Isotonic
• Principal components
• Least angle
• Local
• Segmented
• Errors-in-variables
Estimation
• Least squares
• Ordinary least squares
• Linear (math)
• Partial
• Total
• Generalized
• Weighted
• Non-linear
• Iteratively reweighted
• Ridge regression
• LASSO
• Least absolute deviations
• Bayesian
• Bayesian multivariate
Background
• Regression model validation
• Mean and predicted response
• Errors and residuals
• Goodness of fit
• Studentized residual
• Gauss–Markov theorem
•  Statistics portal
• v
• t
• $e^{[1]}$

In statistics, **regression analysis** is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'Criterion Variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the

average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the **regression function**. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not imply causation.

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results.^{[2][3]}

History

The earliest form of regression was the method of least squares, which was published by Legendre in 1805,^[4] and by Gauss in 1809.^[5] Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821,^[6] including a version of the Gauss–Markov theorem.

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean). For Galton, regression had only this biological meaning,^{[7][8]} but his work was later extended by Udney Yule and Karl Pearson to a more general statistical context. In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation of 1821.

In the 1950s and 1960s, economists used electromechanical desk calculators to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression.^[9]

Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression, regression involving correlated responses such as time series and growth curves, regression in which the predictor or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

Regression models

Regression models involve the following variables:

- The **unknown parameters**, denoted as $\boldsymbol{\beta}$, which may represent a scalar or a vector.
- The **independent variables**, \mathbf{X} .
- The **dependent variable**, Y .

In various fields of application, different terminologies are used in place of dependent and independent variables.

A regression model relates Y to a function of \mathbf{X} and $\boldsymbol{\beta}$.

$$Y \approx f(\mathbf{X}, \boldsymbol{\beta})$$

The approximation is usually formalized as $E(Y|\mathbf{X}) = f(\mathbf{X}, \boldsymbol{\beta})$. To carry out regression analysis, the form of the function f must be specified. Sometimes the form of this function is based on knowledge about the relationship between Y and \mathbf{X} that does not rely on the data. If no such knowledge is available, a flexible or convenient form for f is chosen.

Assume now that the vector of unknown parameters $\boldsymbol{\beta}$ is of length k . In order to perform a regression analysis the user must provide information about the dependent variable Y :

- If N data points of the form (Y, \mathbf{X}) are observed, where $N < k$, most classical approaches to regression analysis cannot be performed: since the system of equations defining the regression model is underdetermined, there are not enough data to recover $\boldsymbol{\beta}$.
- If exactly $N = k$ data points are observed, and the function f is linear, the equations $Y = f(\mathbf{X}, \boldsymbol{\beta})$ can be solved exactly rather than approximately. This reduces to solving a set of N equations with N unknowns (the elements of $\boldsymbol{\beta}$), which has a unique solution as long as the \mathbf{X} are linearly independent. If f is nonlinear, a solution may not exist, or many solutions may exist.
- The most common situation is where $N > k$ data points are observed. In this case, there is enough information in the data to estimate a unique value for $\boldsymbol{\beta}$ that best fits the data in some sense, and the regression model when applied to the data can be viewed as an overdetermined system in $\boldsymbol{\beta}$.

In the last case, the regression analysis provides the tools for:

1. Finding a solution for unknown parameters $\boldsymbol{\beta}$ that will, for example, minimize the distance between the measured and predicted values of the dependent variable Y (also known as method of least squares).
2. Under certain statistical assumptions, the regression analysis uses the surplus of information to provide statistical information about the unknown parameters $\boldsymbol{\beta}$ and predicted values of the dependent variable Y .

Necessary number of independent measurements

Consider a regression model which has three unknown parameters, β_0 , β_1 , and β_2 . Suppose an experimenter performs 10 measurements all at exactly the same value of independent variable vector \mathbf{X} (which contains the independent variables X_1 , X_2 , and X_3). In this case, regression analysis fails to give a unique set of estimated values for the three unknown parameters; the experimenter did not provide enough information. The best one can do is to estimate the average value and the standard deviation of the dependent variable Y . Similarly, measuring at two different values of \mathbf{X} would give enough data for a regression with two unknowns, but not for three or more unknowns.

If the experimenter had performed measurements at three different values of the independent variable vector \mathbf{X} , then regression analysis would provide a unique set of estimates for the three unknown parameters in $\boldsymbol{\beta}$.

In the case of general linear regression, the above statement is equivalent to the requirement that the matrix $\mathbf{X}^T \mathbf{X}$ is invertible.

Statistical assumptions

When the number of measurements, N , is larger than the number of unknown parameters, k , and the measurement errors ϵ_i are normally distributed then *the excess of information* contained in $(N - k)$ measurements is used to make statistical predictions about the unknown parameters. This excess of information is referred to as the degrees of freedom of the regression.

Underlying assumptions

Classical assumptions for regression analysis include:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error. (Note: If this is not so, modeling may be done instead using errors-in-variables model techniques).
- The predictors are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might instead be used.

These are sufficient conditions for the least-squares estimator to possess desirable properties; in particular, these assumptions imply that the parameter estimates will be unbiased, consistent, and efficient in the class of linear unbiased estimators. It is important to note that actual data rarely satisfies the assumptions. That is, the method is used even though the assumptions are not true. Variation from the assumptions can sometimes be used as a measure of how far the model is from being useful. Many of these assumptions may be relaxed in more advanced treatments. Reports of statistical analyses usually include analyses of tests on the sample data and methodology for the fit and usefulness of the model.

Assumptions include the geometrical support of the variables.^[10]Wikipedia:Please clarify Independent and dependent variables often refer to values measured at point locations. There may be spatial trends and spatial autocorrelation in the variables that violate statistical assumptions of regression. Geographic weighted regression is one technique to deal with such data. Also, variables may include values aggregated by areas. With aggregated data the modifiable areal unit problem can cause extreme variation in regression parameters. When analyzing data aggregated by political boundaries, postal codes or census areas results may be very distinct with a different choice of units.

Linear regression

In linear regression, the model specification is that the dependent variable, y_i is a linear combination of the *parameters* (but need not be linear in the *independent variables*). For example, in simple linear regression for modeling n data points there is one independent variable: x_i , and two parameters, β_0 and β_1 :

$$\text{straight line: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n.$$

In multiple linear regression, there are several independent variables or functions of independent variables.

Adding a term in x_i^2 to the preceding regression gives:

$$\text{parabola: } y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad i = 1, \dots, n.$$

This is still linear regression; although the expression on the right hand side is quadratic in the independent variable x_i , it is linear in the parameters β_0 , β_1 and β_2 .

In both cases, ϵ_i is an error term and the subscript i indexes a particular observation.

Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

The residual, $e_i = y_i - \hat{y}_i$, is the difference between the value of the dependent variable predicted by the model, \hat{y}_i , and the true value of the dependent variable, y_i . One method of estimation is ordinary least squares. This method obtains parameter estimates that minimize the sum of squared residuals, SSE,^{[11][12]} also sometimes denoted RSS:

$$SSE = \sum_{i=1}^n e_i^2.$$

Minimization of this function results in a set of normal equations, a set of simultaneous linear equations in the parameters, which are solved to yield the parameter estimators, $\hat{\beta}_0, \hat{\beta}_1$.

In the case of simple regression, the formulas for the least squares estimates are

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where \bar{x} is the mean (average) of the x values and \bar{y} is the mean of the y values.

Under the assumption that the population error term has a constant variance, the estimate of that variance is given by:

$$\hat{\sigma}_\varepsilon^2 = \frac{SSE}{n-2}.$$

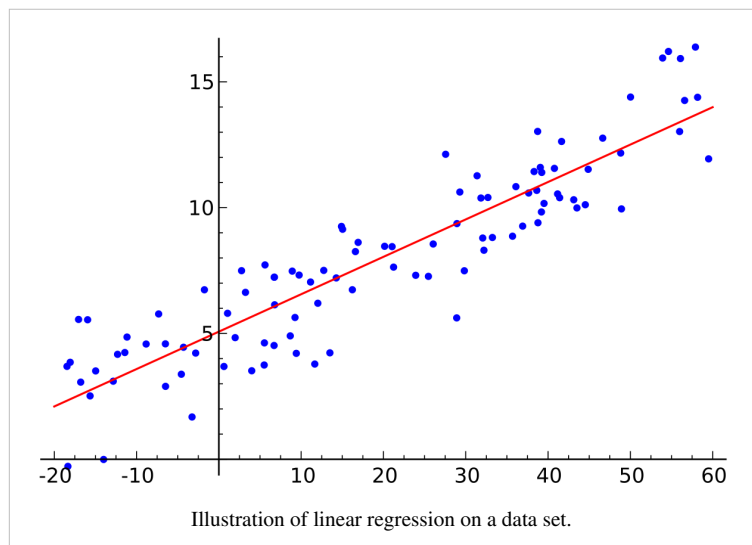
This is called the mean square error (MSE) of the regression. The denominator is the sample size reduced by the number of model parameters estimated from the same data, $(n-p)$ for p regressors or $(n-p-1)$ if an intercept is used.^[13] In this case, $p=1$ so the denominator is $n-2$.

The standard errors of the parameter estimates are given by

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

$$\hat{\sigma}_{\beta_1} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}.$$

Under the further assumption that the population error term is normally distributed, the researcher can use these estimated standard errors to create confidence intervals and conduct hypothesis tests about the population parameters.



General linear model

In the more general multiple regression model, there are p independent variables:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

where x_{ij} is the i^{th} observation on the j^{th} independent variable, and where the first independent variable takes the value 1 for all i (so β_1 is the regression intercept).

The least squares parameter estimates are obtained from p normal equations. The residual can be written as

$$\varepsilon_i = y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}.$$

The **normal equations** are

$$\sum_{i=1}^n \sum_{k=1}^p X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^n X_{ij} y_i, \quad j = 1, \dots, p.$$

In matrix notation, the normal equations are written as

$$(\mathbf{X}^T \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{Y},$$

where the ij element of X is x_{ij} , the i element of the column vector Y is y_i , and the j element of $\hat{\boldsymbol{\beta}}$ is $\hat{\beta}_j$. Thus X is $n \times p$, Y is $n \times 1$, and $\hat{\boldsymbol{\beta}}$ is $p \times 1$. The solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

Diagnostics

Once a regression model has been constructed, it may be important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. Commonly used checks of goodness of fit include the R-squared, analyses of the pattern of residuals and hypothesis testing. Statistical significance can be checked by an F-test of the overall fit, followed by t-tests of individual parameters.

Interpretations of these diagnostic tests rest heavily on the model assumptions. Although examination of the residuals can be used to invalidate a model, the results of a t-test or F-test are sometimes more difficult to interpret if the model's assumptions are violated. For example, if the error term does not have a normal distribution, in small samples the estimated parameters will not follow normal distributions and complicate inference. With relatively large samples, however, a central limit theorem can be invoked such that hypothesis testing may proceed using asymptotic approximations.

"Limited dependent" variables

The phrase "limited dependent" is used in econometric statistics for categorical and constrained variables.

The response variable may be non-continuous ("limited" to lie on some subset of the real line). For binary (zero or one) variables, if analysis proceeds with least-squares linear regression, the model is called the linear probability model. Nonlinear models for binary dependent variables include the probit and logit model. The multivariate probit model is a standard method of estimating a joint relationship between several binary dependent variables and some independent variables. For categorical variables with more than two values there is the multinomial logit. For ordinal variables with more than two values, there are the ordered logit and ordered probit models. Censored regression models may be used when the dependent variable is only sometimes observed, and Heckman correction type models may be used when the sample is not randomly selected from the population of interest. An alternative to such procedures is linear regression based on polychoric correlation (or polyserial correlations) between the categorical variables. Such procedures differ in the assumptions made about the distribution of the variables in the population. If the variable is positive with low values and represents the repetition of the occurrence of an event, then count models like the Poisson regression or the negative binomial model may be used instead.

Interpolation and extrapolation

Regression models predict a value of the Y variable given known values of the X variables. Prediction *within* the range of values in the dataset used for model-fitting is known informally as interpolation. Prediction *outside* this range of the data is known as extrapolation. Performing extrapolation relies strongly on the regression assumptions. The further the extrapolation goes outside the data, the more room there is for the model to fail due to differences between the assumptions and the sample data or the true values.

It is generally advised ^[citation needed] that when performing extrapolation, one should accompany the estimated value of the dependent variable with a prediction interval that represents the uncertainty. Such intervals tend to expand rapidly as the values of the independent variable(s) moved outside the range covered by the observed data.

For such reasons and others, some tend to say that it might be unwise to undertake extrapolation. ^[14]

However, this does not cover the full set of modelling errors that may be being made: in particular, the assumption of a particular form for the relation between Y and X . A properly conducted regression analysis will include an assessment of how well the assumed form is matched by the observed data, but it can only do so within the range of values of the independent variables actually available. This means that any extrapolation is particularly reliant on the assumptions being made about the structural form of the regression relationship. Best-practice advice here ^[citation needed] is that a linear-in-variables and linear-in-parameters relationship should not be chosen simply for computational convenience, but that all available knowledge should be deployed in constructing a regression model. If this knowledge includes the fact that the dependent variable cannot go outside a certain range of values, this can be made use of in selecting the model – even if the observed dataset has no values particularly near such bounds. The implications of this step of choosing an appropriate functional form for the regression can be great when extrapolation is considered. At a minimum, it can ensure that any extrapolation arising from a fitted model is "realistic" (or in accord with what is known).

Nonlinear regression

When the model function is not linear in the parameters, the sum of squares must be minimized by an iterative procedure. This introduces many complications which are summarized in Differences between linear and non-linear least squares

Power and sample size calculations

There are no generally agreed methods for relating the number of observations versus the number of independent variables in the model. One rule of thumb suggested by Good and Hardin is $N = m^n$, where N is the sample size, n is the number of independent variables and m is the number of observations needed to reach the desired precision if the model had only one independent variable. For example, a researcher is building a linear regression model using a dataset that contains 1000 patients (N). If he decides that five observations are needed to precisely define a straight line (m), then the maximum number of independent variables his model can support is 4, because

$$\frac{\log 1000}{\log 5} = 4.29.$$

Other methods

Although the parameters of a regression model are usually estimated using the method of least squares, other methods which have been used include:

- Bayesian methods, e.g. Bayesian linear regression
- Percentage regression, for situations where reducing *percentage* errors is deemed more appropriate.
- Least absolute deviations, which is more robust in the presence of outliers, leading to quantile regression
- Nonparametric regression, requires a large number of observations and is computationally intensive
- Distance metric learning, which is learned by the search of a meaningful distance metric in a given input space.

Software

All major statistical software packages perform least squares regression analysis and inference. Simple linear regression and multiple regression using least squares can be done in some spreadsheet applications and on some calculators. While many statistical software packages can perform various types of nonparametric and robust regression, these methods are less standardized; different software packages implement different methods, and a method with a given name may be implemented differently in different packages. Specialized regression software has been developed for use in fields such as survey analysis and neuroimaging.

References

- [1] http://en.wikipedia.org/w/index.php?title=Template:Regression_bar&action=edit
- [2] David A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press (2005)
- [3] R. Dennis Cook; Sanford Weisberg Criticism and Influence Analysis in Regression ([http://links.jstor.org/sici?sici=0081-1750\(1982\)13<313:CAIAIR>2.0.CO;2-3](http://links.jstor.org/sici?sici=0081-1750(1982)13<313:CAIAIR>2.0.CO;2-3)), *Sociological Methodology*, Vol. 13. (1982), pp. 313–361
- [4] A.M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes* (<http://books.google.ca/books?id=FRcOAAAAQAAJ>), Firmin Didot, Paris, 1805. "Sur la Méthode des moindres carrés" appears as an appendix.
- [5] C.F. Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*. (1809)
- [6] C.F. Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae* (http://books.google.co.za/books?id=ZQ8OAAAAQAAJ&printsec=frontcover&dq=Theoria+combinationis+observationum+erroribus+minimis+obnoxiae&as_brr=3#v=onepage&q=&f=false). (1821/1823)
- [7] Francis Galton. "Typical laws of heredity", *Nature* 15 (1877), 492–495, 512–514, 532–533. (*Galton uses the term "reversion" in this paper, which discusses the size of peas.*)
- [8] Francis Galton. Presidential address, Section H, Anthropology. (1885) (*Galton uses the term "regression" in this paper, which discusses the height of humans.*)
- [9] Rodney Ramcharan. Regressions: Why Are Economists Obsessed with Them? (<http://www.imf.org/external/pubs/ft/fandd/2006/03/basics.htm>) March 2006. Accessed 2011-12-03.
- [10] N. Cressie (1996) Change of Support and the Modifiable Areal Unit Problem. *Geographical Systems* 3:159–180.
- [11] M. H. Kutner, C. J. Nachtsheim, and J. Neter (2004), "Applied Linear Regression Models", 4th ed., McGraw-Hill/Irwin, Boston (p. 25)
- [12] N. Ravishanker and D. K. Dey (2002), "A First Course in Linear Model Theory", Chapman and Hall/CRC, Boca Raton (p. 101)
- [13] Steel, R.G.D, and Torrie, J. H., *Principles and Procedures of Statistics with Special Reference to the Biological Sciences.*, McGraw Hill, 1960, page 288.
- [14] Chiang, C.L, (2003) *Statistical methods of analysis*, World Scientific. ISBN 981-238-310-7 - page 274 section 9.7.4 "interpolation vs extrapolation" (<http://books.google.com/books?id=BuPNIbaN5v4C&lpg=PA274&dq=regression+extrapolation&pg=PA274#v=onepage&q=regression+extrapolation&f=false>)

Further reading

- William H. Kruskal and Judith M. Tanur, ed. (1978), "Linear Hypotheses," *International Encyclopedia of Statistics*. Free Press, v. 1,
 - Evan J. Williams, "I. Regression," pp. 523–41.
 - Julian C. Stanley, "II. Analysis of Variance," pp. 541–554.
- Lindley, D.V. (1987). "Regression and correlation analysis," New Palgrave: A Dictionary of Economics, v. 4, pp. 120–23.
- Birkes, David and Dodge, Y., *Alternative Methods of Regression*. ISBN 0-471-56881-3
- Chatfield, C. (1993) "Calculating Interval Forecasts," *Journal of Business and Economic Statistics*, **11**, pp. 121–135.
- Draper, N.R.; Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). John Wiley. ISBN 0-471-17082-8.
- Fox, J. (1997). *Applied Regression Analysis, Linear Models and Related Methods*. Sage
- Hardle, W., *Applied Nonparametric Regression* (1990), ISBN 0-521-42950-1
- Meade, N. and T. Islam (1995) "Prediction Intervals for Growth Curve Forecasts" (<http://onlinelibrary.wiley.com/doi/10.1002/for.3980140502/abstract>) *Journal of Forecasting*, **14**, pp. 413–430.
- A. Sen, M. Srivastava, *Regression Analysis — Theory, Methods, and Applications*, Springer-Verlag, Berlin, 2011 (4th printing).
- T. Strutz: *Data Fitting and Uncertainty (A practical introduction to weighted least squares and beyond)*. Vieweg+Teubner, ISBN 978-3-8348-1022-9.

External links

- Hazewinkel, Michiel, ed. (2001), "Regression analysis" (<http://www.encyclopediaofmath.org/index.php?title=p/r080620>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Earliest Uses: Regression (<http://jeff560.tripod.com/r.html>) – basic history and references
- Regression of Weakly Correlated Data (http://www.vias.org/simulations/simusoftware_regrot.html) – how linear regression mistakes can appear when Y-range is much smaller than X-range
- Statistical interpolation with ordinary least squares (<http://commons.wikimedia.org/wiki/File:Public-domain-ordinary-least-squares.gif>)

Path analysis (statistics)

In statistics, **path analysis** is used to describe the directed dependencies among a set of variables. This includes models equivalent to any form of multiple regression analysis, factor analysis, canonical correlation analysis, discriminant analysis, as well as more general families of models in the multivariate analysis of variance and covariance analyses (MANOVA, ANOVA, ANCOVA).

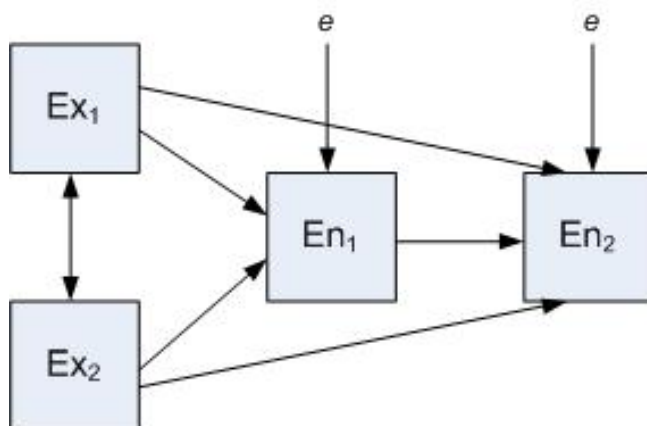
In addition to being thought of as a form of multiple regression focusing on causality, path analysis can be viewed as a special case of structural equation modeling (SEM) – one in which only single indicators are employed for each of the variables in the causal model. That is, path analysis is SEM with a structural model, but no measurement model. Other terms used to refer to path analysis include causal modeling, analysis of covariance structures, and latent variable models.

History

Path analysis was developed around 1918 by geneticist Sewall Wright, who wrote about it more extensively in the 1920s. It has since been applied to a vast array of complex modeling areas, including biology, sociology, and econometrics.^[1]

Path modeling

In the model below, the two exogenous variables (Ex_1 and Ex_2) are modeled as being correlated and as having both direct and indirect (through En_1) effects on En_2 (the two dependent or 'endogenous' variables). In most real models, the endogenous variables are also affected by factors outside the model (including measurement error). The effects of such extraneous variables are depicted by the "e" or error terms in the model.



Using the same variables, alternative models are conceivable. For example, it may be hypothesized that Ex_1 has only an indirect effect on En_2 , deleting the arrow from Ex_1 to En_2 ; and the likelihood or 'fit' of these two models can be compared statistically.

Path tracing rules

In order to validly calculate the relationship between any two boxes in the diagram, Wright (1934) proposed a simple set of path tracing rules, for calculating the correlation between two variables. The correlation is equal to the sum of the contribution of all the pathways through which the two variables are connected. The strength of each of these contributing pathways is calculated as the product of the path-coefficients along that pathway.

The rules for path tracing are:

1. You can trace backward up an arrow and then forward along the next, or forwards from one variable to the other, but never forward and then back.
2. You can pass through each variable only once in a given chain of paths.
3. No more than one bi-directional arrow can be included in each path-chain.

Another way to think of rule one is that you can never pass out of one arrow head and into another arrowhead: heads-tails, or tails-heads, not heads-heads.

Again, the expected correlation due to each chain traced between two variables is the product of the standardized path coefficients, and the total expected correlation between two variables is the sum of these contributing path-chains.

NB: Wright's rules assume a model without feedback loops: the directed graph of the model must contain no cycles.

Path tracing in unstandardized models

If the modeled variables have not been standardized, an additional rule allows the expected covariances to be calculated as long as no paths exist connecting dependent variables to other dependent variables.

The simplest case obtains where all residual variances are modeled explicitly. In this case, in addition to the three rules above, calculate expected covariances by:

1. Compute the product of coefficients in each route between the variables of interest, tracing backwards, changing direction at a two-headed arrow, then tracing forwards.
2. Sum over all distinct routes, where pathways are considered distinct if they contain different coefficients, or encounter those coefficients in a different order.

Where residual variances are not explicitly included, or as a more general solution, at any change of direction encountered in a route (except for at two-way arrows), include the variance of the variable at the point of change. That is, in tracing a path from a dependent variable to an independent variable, include the variance of the independent-variable except where so doing would violate rule 1 above (passing through adjacent arrowheads: i.e., when the independent variable also connects to a double-headed arrow connecting it to another independent variable). In deriving variances (which is necessary in the case where they are not modeled explicitly), the path from a dependent variable into an independent variable and back is counted once only.

References

- [1] Dodge, Y. (2003) *The Oxford Dictionary of Statistical Terms*. OUP. ISBN 0-19-920613-9

Analysis

Moving average

In statistics, a **moving average** (**rolling average** or **running average**) is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. It is also called a **moving mean (MM)**^[1] or **rolling mean** and is a type of finite impulse response filter. Variations include: simple, and cumulative, or weighted forms (described below).

Given a series of numbers and a fixed subset size, the first element of the moving average is obtained by taking the average of the initial fixed subset of the number series. Then the subset is modified by "shifting forward"; that is, excluding the first number of the series and including the next number following the original subset in the series. This creates a new subset of numbers, which is averaged. This process is repeated over the entire data series. The plot line connecting all the (fixed) averages is the moving average. A moving average is a set of numbers, each of which is the average of the corresponding subset of a larger set of datum points. A moving average may also use unequal weights for each datum value in the subset to emphasize particular values in the subset.

A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly. For example, it is often used in technical analysis of financial data, like stock prices, returns or trading volumes. It is also used in economics to examine gross domestic product, employment or other macroeconomic time series. Mathematically, a moving average is a type of convolution and so it can be viewed as an example of a low-pass filter used in signal processing. When used with non-time series data, a moving average filters higher frequency components without any specific connection to time, although typically some kind of ordering is implied. Viewed simplistically it can be regarded as smoothing the data.



Simple moving average

In financial applications a **simple moving average (SMA)** is the unweighted mean of the previous n datum points. However, in science and engineering the mean is normally taken from an equal number of data on either side of a central value. This ensures that variations in the mean are aligned with the variations in the data rather than being shifted in time. An example of a simple equally weighted running mean for a n -day sample of closing price is the mean of the previous n days' closing prices. If those prices are $P_M, P_{M-1}, \dots, P_{M-(n-1)}$ then the formula is

$$SMA = \frac{P_M + P_{M-1} + \dots + P_{M-(n-1)}}{n}$$

When calculating successive values, a new value comes into the sum and an old value drops out, meaning a full summation each time is unnecessary for this simple case,



$$SMA_{\text{today}} = SMA_{\text{yesterday}} - \frac{P_{M-n}}{n} + \frac{P_M}{n}$$

The period selected depends on the type of movement of interest, such as short, intermediate, or long-term. In financial terms moving-average levels can be interpreted as support in a rising market, or resistance in a falling market.

If the data used are not centered around the mean, a simple moving average lags behind the latest datum point by half the sample width. An SMA can also be disproportionately influenced by old datum points dropping out or new data coming in. One characteristic of the SMA is that if the data have a periodic fluctuation, then applying an SMA of that period will eliminate that variation (the average always containing one complete cycle). But a perfectly regular cycle is rarely encountered.^[2]

For a number of applications, it is advantageous to avoid the shifting induced by using only 'past' data. Hence a **central moving average** can be computed, using data equally spaced on either side of the point in the series where the mean is calculated. This requires using an odd number of datum points in the sample window.

A major drawback of the SMA is that it lets through a significant amount of the signal shorter than the window length. Worse, it **actually inverts it**. This can lead to unexpected artifacts, such as peaks in the "smoothed" result appearing where there were troughs in the data. It also leads to the result being less "smooth" than expected since some of the higher frequencies are not properly removed.

The problem can be over-come by repeating the process three times with the window being shortened by a factor of 1.4303 at each step.^[3] This removes the negation effects and provides a well-behaved filter. This solution is often used in real-time audio filtering since it is computationally quicker than other comparable filters such as a gaussian kernel.

An example of inversion defect in SMA and the application of repeating SMA to avoid it can be illustrated here: <http://www.woodfortrees.org/plot/rss/from:1980/plot/rss/from:1980/mean:60/plot/rss/from:1980/mean:30/mean:22/mean:17>

Cumulative moving average

In a **cumulative moving average**, the data arrive in an ordered datum stream, and the user would like to get the average of all of the data up until the current datum point. For example, an investor may want the average price of all of the stock transactions for a particular stock up until the current time. As each new transaction occurs, the average price at the time of the transaction can be calculated for all of the transactions up to that point using the cumulative average, typically an equally weighted average of the sequence of i values x_1, \dots, x_i up to the current time:

$$CA_i = \frac{x_1 + \dots + x_i}{i}.$$

The brute-force method to calculate this would be to store all of the data and calculate the sum and divide by the number of datum points every time a new datum point arrived. However, it is possible to simply update cumulative average as a new value, x_{i+1} becomes available, using the formula:

$$CA_{i+1} = \frac{x_{i+1} + iCA_i}{i+1},$$

where CA_0 can be taken to be equal to 0.

Thus the current cumulative average for a new datum point is equal to the previous cumulative average, times i , plus the latest datum point, all divided by the number of points received so far, $i+1$. When all of the datum points arrive ($i = N$), then the cumulative average will equal the final average.

The derivation of the cumulative average formula is straightforward. Using

$$x_1 + \dots + x_i = iCA_i,$$

and similarly for $i+1$, it is seen that

$$x_{i+1} = (x_1 + \dots + x_{i+1}) - (x_1 + \dots + x_i) = (i + 1)CA_{i+1} - iCA_i.$$

Solving this equation for CA_{i+1} results in:

$$CA_{i+1} = \frac{(x_{i+1} + iCA_i)}{i + 1} = CA_i + \frac{x_{i+1} - CA_i}{i + 1}.$$

Weighted moving average

A weighted average is any average that has multiplying factors to give different weights to data at different positions in the sample window. Mathematically, the moving average is the convolution of the datum points with a fixed weighting function. One application is removing pixelisation from a digital graphical image.

In technical analysis of financial data, a **weighted moving average** (WMA) has the specific meaning of weights that decrease in arithmetical progression. In an n -day WMA the latest day has weight n , the second latest $n - 1$, etc., down to one.

$$WMA_M = \frac{np_M + (n - 1)p_{M-1} + \dots + 2p_{(M-n+2)} + p_{(M-n+1)}}{n + (n - 1) + \dots + 2 + 1}$$

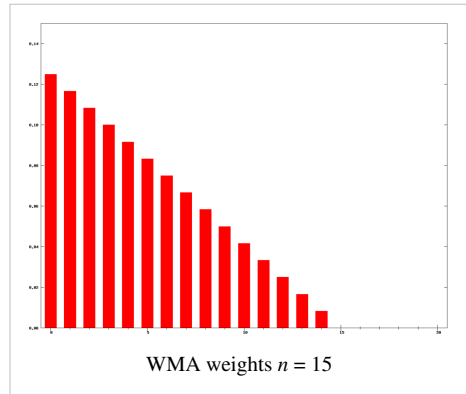
The denominator is a triangle number equal to $\frac{n(n + 1)}{2}$. In the more general case the denominator will always be the sum of the individual weights.

When calculating the WMA across successive values, the difference between the numerators of WMA_{M+1} and WMA_M is $np_{M+1} - p_M - \dots - p_{M-n+1}$. If we denote the sum $p_M + \dots + p_{M-n+1}$ by $Total_M$, then

$$Total_{M+1} = Total_M + p_{M+1} - p_{M-n+1}$$

$$Numerator_{M+1} = Numerator_M + np_{M+1} - Total_M$$

$$WMA_{M+1} = \frac{Numerator_{M+1}}{n + (n - 1) + \dots + 2 + 1}$$



The graph at the right shows how the weights decrease, from highest weight for the most recent datum points, down to zero. It can be compared to the weights in the exponential moving average which follows.

Exponential moving average

An **exponential moving average** (EMA), also known as an **exponentially weighted moving average** (EWMA),^[4] is a type of infinite impulse response filter that applies weighting factors which decrease exponentially. The weighting for each older datum decreases exponentially, never reaching zero. The graph at right shows an example of the weight decrease.

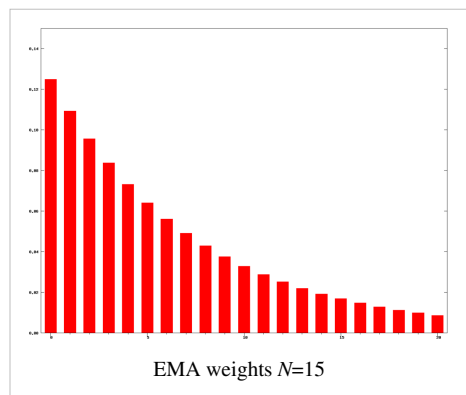
The EMA for a series Y may be calculated recursively:

$$S_1 = Y_1$$

$$\text{for } t > 1, \quad S_t = \alpha \cdot Y_{t-1} + (1 - \alpha) \cdot S_{t-1}$$

Where:

- The coefficient α represents the degree of weighting decrease, a constant smoothing factor between 0 and 1. A higher α discounts older observations faster.
- Y_t is the value at a time period t .



- S_t is the value of the EMA at any time period t .

S_1 is undefined. S_1 may be initialized in a number of different ways, most commonly by setting S_1 to Y_1 , though other techniques exist, such as setting S_1 to an average of the first 4 or 5 observations. The importance of the S_1 initialisations effect on the resultant moving average depends on α ; smaller α values make the choice of S_1 relatively more important than larger α values, since a higher α discounts older observations faster.

Whatever is done for S_1 it assumes something about values prior to the available data and is necessarily in error. In view of this the early results should be regarded as unreliable until the iterations have had time to converge. This is sometimes called a 'spin-up' interval. One way to assess when it can be regarded as reliable is consider the required accuracy of the result. For example, if 3% accuracy is required, initialising with Y_1 and taking data after five time constants (defined above) will ensure that the calculation has converged to within 3% (only <3% of Y_1 will remain in the result). Sometimes with very small alpha, this can mean little of the result is useful. This is analogous to the problem of using a convolution filter (such as a weighted average) with a very long window.

This formulation is according to Hunter (1986).^[5] By repeated application of this formula for different times, we can eventually write S_t as a weighted sum of the datum points Y_t , as:

$$S_t = \alpha \times (Y_{t-1} + (1-\alpha) \times Y_{t-2} + (1-\alpha)^2 \times Y_{t-3} + \dots + (1-\alpha)^k \times Y_{t-(k+1)}) + (1-\alpha)^{k+1} \times S_{t-(k+1)}$$

for any suitable $k = 0, 1, 2, \dots$. The weight of the general datum point Y_{t-i} is $\alpha(1 - \alpha)^{i-1}$.

An alternate approach by Roberts (1959) uses Y_t in lieu of Y_{t-1} :^[6]

$$S_{t, \text{alternate}} = \alpha \cdot Y_t + (1 - \alpha) \cdot S_{t-1}$$

This formula can also be expressed in technical analysis terms as follows, showing how the EMA steps towards the latest datum point, but only by a proportion of the difference (each time):

$$EMA_{\text{today}} = EMA_{\text{yesterday}} + \alpha \times (\text{price}_{\text{today}} - EMA_{\text{yesterday}})$$

Expanding out $EMA_{\text{yesterday}}$ each time results in the following power series, showing how the weighting factor on each datum point p_1, p_2 , etc., decreases exponentially:

$$EMA_{\text{today}} = \alpha \times (p_1 + (1 - \alpha)p_2 + (1 - \alpha)^2 p_3 + (1 - \alpha)^3 p_4 + \dots)$$

where

- p_1 is $\text{price}_{\text{today}}$
- p_2 is $\text{price}_{\text{yesterday}}$
- and so on

$$EMA_{\text{today}} = \frac{p_1 + (1 - \alpha)p_2 + (1 - \alpha)^2 p_3 + (1 - \alpha)^3 p_4 + \dots}{1 + (1 - \alpha) + (1 - \alpha)^2 + (1 - \alpha)^3 + \dots},$$

since $1/\alpha = 1 + (1 - \alpha) + (1 - \alpha)^2 + \dots$.

This is an infinite sum with decreasing terms.

The N periods in an N -day EMA only specify the α factor. N is not a stopping point for the calculation in the way it is in an SMA or WMA. For sufficiently large N , the first N datum points in an EMA represent about 86% of the total weight in the calculation when $\alpha = 2/(N + 1)$:^[7]

$$\frac{\alpha \times (1 + (1 - \alpha) + (1 - \alpha)^2 + \dots + (1 - \alpha)^N)}{\alpha \times (1 + (1 - \alpha) + (1 - \alpha)^2 + \dots + (1 - \alpha)^\infty)} = 1 - \left(1 - \frac{2}{N + 1}\right)^{N+1}$$

i.e. $\lim_{N \rightarrow \infty} \left[1 - \left(1 - \frac{2}{N + 1}\right)^{N+1}\right]$ simplified,^[8] tends to $1 - e^{-2} \approx 0.8647$.

The above discussion requires a bit of clarification. The sum of the weights of all the terms (i.e., infinite number of terms) in an exponential moving average is 1. The sum of the weights of N terms is $1 - (1 - \alpha)^{N+1}$. Both of these sums can be derived by using the formula for the sum of a geometric series. The weight omitted after N terms is given by subtracting this from 1, and you get $1 - (1 - (1 - \alpha)^{N+1}) = (1 - \alpha)^{N+1}$ (this is essentially the

formula given below for the weight omitted). Note that there is no "accepted" value that should be chosen for α although there are some recommended values based on the application. In the above discussion, we have substituted a commonly used value for $\alpha = 2/(N + 1)$ in the formula for the weight of N terms. This value for α comes from setting the average of the data from a SMA equal to the average age of the data from an EWA and solving for α . Again, it is just a recommendation—not a requirement. If you make this substitution, and you make use of ^[9] $\lim_{n \rightarrow \infty} \left(1 + \frac{a}{1+n}\right)^n = e^a$, then you have an approximation. Intuitively, what this is telling us is that the weight after N terms of an "N-period" exponential moving average converges to 0.864.

The power formula above gives a starting value for a particular day, after which the successive days formula shown first can be applied. The question of how far back to go for an initial value depends, in the worst case, on the data. Large price values in old data will affect on the total even if their weighting is very small. If prices have small variations then just the weighting can be considered. The weight omitted by stopping after k terms is

$$\alpha \times \left((1 - \alpha)^k + (1 - \alpha)^{k+1} + (1 - \alpha)^{k+2} + \dots \right),$$

which is

$$\alpha \times (1 - \alpha)^k \times \left(1 + (1 - \alpha) + (1 - \alpha)^2 + \dots \right),$$

i.e. a fraction

$$\begin{aligned} \frac{\text{weight omitted by stopping after k terms}}{\text{total weight}} &= \frac{\alpha \times \left[(1 - \alpha)^k + (1 - \alpha)^{k+1} + (1 - \alpha)^{k+2} + \dots \right]}{\alpha \times \left[1 + (1 - \alpha) + (1 - \alpha)^2 + \dots \right]} \\ &= \frac{\alpha(1 - \alpha)^k \times \frac{1}{1 - (1 - \alpha)}}{\frac{\alpha}{1 - (1 - \alpha)}} \\ &= (1 - \alpha)^k \end{aligned}$$

out of the total weight.

For example, to have 99.9% of the weight, set above ratio equal to 0.1% and solve for k :

$$k = \frac{\log(0.001)}{\log(1 - \alpha)}$$

terms should be used. Since $\log(1 - \alpha)$ approaches $\frac{-2}{N + 1}$ as N increases,^[10] this simplifies to approximately^[11]

$$k = 3.45(N + 1)$$

for this example (99.9% weight).

Modified moving average

A **modified moving average** (MMA), **running moving average** (RMA), or **smoothed moving average** is defined as:

$$\text{MMA}_{\text{today}} = \frac{(N - 1) \times \text{MMA}_{\text{yesterday}} + \text{price}}{N}$$

In short, this is an exponential moving average, with $\alpha = 1/N$.

Application to measuring computer performance

Some computer performance metrics, e.g. the average process queue length, or the average CPU utilization, use a form of exponential moving average.

$$S_n = \alpha(t_n - t_{n-1}) \times Y_n + (1 - \alpha(t_n - t_{n-1})) \times S_{n-1}.$$

Here α is defined as a function of time between two readings. An example of a coefficient giving bigger weight to the current reading, and smaller weight to the older readings is

$$\alpha(t_n - t_{n-1}) = 1 - \exp\left(-\frac{t_n - t_{n-1}}{W \times 60}\right)$$

where $\exp()$ is the exponential function, time for readings t_n is expressed in seconds, and W is the period of time in minutes over which the reading is said to be averaged (the mean lifetime of each reading in the average). Given the above definition of α , the moving average can be expressed as

$$S_n = \left(1 - \exp\left(-\frac{t_n - t_{n-1}}{W \times 60}\right)\right) \times Y_n + \exp\left(-\frac{t_n - t_{n-1}}{W \times 60}\right) \times S_{n-1}$$

For example, a 15-minute average L of a process queue length Q , measured every 5 seconds (time difference is 5 seconds), is computed as

$$L_n = \left(1 - \exp\left(-\frac{5}{15 \times 60}\right)\right) \times Q_n + e^{-\frac{5}{15 \times 60}} \times L_{n-1} = \left(1 - \exp\left(-\frac{1}{180}\right)\right) \times Q_n + e^{-1/180} \times L_{n-1} = Q_n + e^{-1/180} \times (L_{n-1} - Q_n)$$

Other weightings

Other weighting systems are used occasionally – for example, in share trading a **volume weighting** will weight each time period in proportion to its trading volume.

A further weighting, used by actuaries, is Spencer's 15-Point Moving Average^[12] (a central moving average). The symmetric weight coefficients are $-3, -6, -5, 3, 21, 46, 67, 74, 67, 46, 21, 3, -5, -6, -3$.

Outside the world of finance, weighted running means have many forms and applications. Each weighting function or "kernel" has its own characteristics. In engineering and science the frequency and phase response of the filter is often of primary importance in understanding the desired and undesired distortions that a particular filter will apply to the data.

A mean does not just "smooth" the data. A mean is a form of low-pass filter. The effects of the particular filter used should be understood in order to make an appropriate choice. On this point, the French version of this article discusses the spectral effects of 3 kinds of means (cumulative, exponential, Gaussian).

Moving median

From a statistical point of view, the moving average, when used to estimate the underlying trend in a time series, is susceptible to rare events such as rapid shocks or other anomalies. A more robust estimate of the trend is the **simple moving median** over n time points:

$$SMM = \text{Median}(p_M, p_{M-1}, \dots, p_{M-n+1})$$

where the median is found by, for example, sorting the values inside the brackets and finding the value in the middle. For larger values of n , the median can be efficiently computed by updating an indexable skiplist.^[13]

Statistically, the moving average is optimal for recovering the underlying trend of the time series when the fluctuations about the trend are normally distributed. However, the normal distribution does not place high probability on very large deviations from the trend which explains why such deviations will have a disproportionately large effect on the trend estimate. It can be shown that if the fluctuations are instead assumed to be Laplace distributed, then the moving median is statistically optimal.^[14] For a given variance, the Laplace distribution places higher probability on rare events than does the normal, which explains why the moving median tolerates shocks better than the moving mean.

When the simple moving median above is central, the smoothing is identical to the median filter which has applications in, for example, image signal processing.

Notes and references

- [1] Hydrologic Variability of the Cosumnes River Floodplain (http://www.waterboards.ca.gov/waterrights/water_issues/programs/bay_delta/docs/cmmt091412/sldmwa/booth_et_al_2006.pdf) (Booth et. al., San Francisco Estuary and Watershed Science, Volume 4, Issue 2, 2006)
- [2] *Statistical Analysis*, Ya-lun Chou, Holt International, 1975, ISBN 0-03-089422-0, section 17.9.
- [3] <http://climategrog.wordpress.com/2013/05/19/triple-running-mean-filters/>
- [4] <http://lorien.ncl.ac.uk/ming/filter/filewma.htm>
- [5] NIST/SEMATECH e-Handbook of Statistical Methods: Single Exponential Smoothing (<http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc431.htm>) at the National Institute of Standards and Technology
- [6] NIST/SEMATECH e-Handbook of Statistical Methods: EWMA Control Charts (<http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc324.htm>) at the National Institute of Standards and Technology
- [7] The denominator on the left-hand side should be unity, and the numerator will become the right-hand side (geometric series), UNIQ-math-0-efccba492c5c7a1c-QINU .
- [8] Because $(1+x/n)^n$ tends to the limit e^x for large n .
- [9] See the following link (<http://options-trading-notes.blogspot.com/2013/06/derivation-of-ea.html>) for a proof.
- [10] It means $P_M, P_{M-1}, \dots, P_{M-(n-1)} \rightarrow 0$, and the Taylor series of $SMA = \frac{P_M + P_{M-1} + \dots + P_{M-(n-1)}}{n}$ is equivalent to $SMA_{today} = SMA_{yesterday} - \frac{P_{M-n}}{n} + \frac{P_M}{n}$.
- [11] $\log_e(0.001) / 2 = -3.45$
- [12] Spencer's 15-Point Moving Average — from Wolfram MathWorld (<http://mathworld.wolfram.com/Spencers15-PointMovingAverage.html>)
- [13] <http://code.activestate.com/recipes/576930/>
- [14] G.R. Arce, "Nonlinear Signal Processing: A Statistical Approach", Wiley:New Jersey, USA, 2005.

Student's t-test

A **t-test** is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's t distribution.

History

The t -statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name).^{[1][2]} Gosset had been hired due to Claude Guinness's policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness's industrial processes. Gosset devised the t -test as a cheap way to monitor the quality of stout. The student t -test work was submitted to and accepted in the journal *Biometrika*, the journal that Karl Pearson had co-founded and was the Editor-in-Chief; the article was published in 1908. Company policy at Guinness forbade its chemists from publishing their findings, so Gosset published his mathematical work under the pseudonym "Student". Actually, Guinness had a policy of allowing technical staff leave for study (so-called study leave), which Gosset used during the first two terms of the 1906-1907 academic year in Professor Karl Pearson's Biometric Laboratory at University College London. Gosset's identity was then known to fellow statisticians and the Editor-in-Chief Karl Pearson. It is not clear how much of the work Gosset performed while he was at Guinness and how much was done when he was on study leave at University College London.

Uses

Among the most frequently used *t*-tests are:

- A one-sample location test of whether the mean of a population has a value specified in a null hypothesis.
- A two-sample location test of the null hypothesis that the means of two populations are equal. All such tests are usually called **Student's *t*-tests**, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal; the form of the test used when this assumption is dropped is sometimes called Welch's *t*-test. These tests are often referred to as "unpaired" or "independent samples" *t*-tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping.
- A test of the null hypothesis that the difference between two responses measured on the same statistical unit has a mean value of zero. For example, suppose we measure the size of a cancer patient's tumor before and after a treatment. If the treatment is effective, we expect the tumor size for many of the patients to be smaller following the treatment. This is often referred to as the "paired" or "repeated measures" *t*-test: see paired difference test.
- A test of whether the slope of a regression line differs significantly from 0.

Assumptions

Most *t*-test statistics have the form $t = Z/s$, where Z and s are functions of the data. Typically, Z is designed to be sensitive to the alternative hypothesis (i.e., its magnitude tends to be larger when the alternative hypothesis is true), whereas s is a scaling parameter that allows the distribution of t to be determined.

As an example, in the one-sample *t*-test $Z = \bar{X}/(\hat{\sigma}/\sqrt{n})$, where \bar{X} is the sample mean of the data, n is the sample size, and $\hat{\sigma}$ is the population standard deviation of the data. s is the sample standard deviation.

The assumptions underlying a *t*-test are that

- Z follows a standard normal distribution under the null hypothesis
- s^2 follows a χ^2 distribution with p degrees of freedom under the null hypothesis, where p is a positive constant
- Z and s are independent.

In a specific type of *t*-test, these conditions are consequences of the population being studied, and of the way in which the data are sampled. For example, in the *t*-test comparing the means of two independent samples, the following assumptions should be met:

- Each of the two populations being compared should follow a normal distribution. This can be tested using a normality test, such as the Shapiro–Wilk or Kolmogorov–Smirnov test, or it can be assessed graphically using a normal quantile plot.
- If using Student's original definition of the *t*-test, the two populations being compared should have the same variance (testable using F test, Levene's test, Bartlett's test, or the Brown–Forsythe test; or assessable graphically using a Q–Q plot). If the sample sizes in the two groups being compared are equal, Student's original *t*-test is highly robust to the presence of unequal variances. Welch's *t*-test is insensitive to equality of the variances regardless of whether the sample sizes are similar.
- The data used to carry out the test should be sampled independently from the two populations being compared. This is in general not testable from the data, but if the data are known to be dependently sampled (i.e. if they were sampled in clusters), then the classical *t*-tests discussed here may give misleading results.

Unpaired and paired two-sample *t*-tests

Two-sample *t*-tests for a difference in mean involve independent samples, paired samples and overlapping samples. Paired *t*-tests are a form of blocking, and have greater power than unpaired tests when the paired units are similar with respect to "noise factors" that are independent of membership in the two groups being compared.^[3] In a different context, paired *t*-tests can be used to reduce the effects of confounding factors in an observational study.

Independent (unpaired) samples

The independent samples *t*-test is used when two separate sets of independent and identically distributed samples are obtained, one from each of the two populations being compared. For example, suppose we are evaluating the effect of a medical treatment, and we enroll 100 subjects into our study, then randomly assign 50 subjects to the treatment group and 50 subjects to the control group. In this case, we have two independent samples and would use the unpaired form of the *t*-test. The randomization is not essential here – if we contacted 100 people by phone and obtained each person's age and gender, and then used a two-sample *t*-test to see whether the mean ages differ by gender, this would also be an independent samples *t*-test, even though the data are observational.

Paired samples

Paired samples *t*-tests typically consist of a sample of matched pairs of similar units, or one group of units that has been tested twice (a "repeated measures" *t*-test).

A typical example of the repeated measures *t*-test would be where subjects are tested prior to a treatment, say for high blood pressure, and the same subjects are tested again after treatment with a blood-pressure lowering medication. By comparing the same patient's numbers before and after treatment, we are effectively using each patient as their own control. That way the correct rejection of the null hypothesis (here: of no difference made by the treatment) can become much more likely, with statistical power increasing simply because the random between-patient variation has now been eliminated. Note however that an increase of statistical power comes at a price: more tests are required, each subject having to be tested twice. Because half of the sample now depends on the other half, the paired version of Student's *t*-test has only ' $n/2 - 1$ ' degrees of freedom (with ' n ' being the total number of observations). Pairs become individual test units, and the sample has to be doubled to achieve the same number of degrees of freedom.

A paired samples *t*-test based on a "matched-pairs sample" results from an unpaired sample that is subsequently used to form a paired sample, by using additional variables that were measured along with the variable of interest. The matching is carried out by identifying pairs of values consisting of one observation from each of the two samples, where the pair is similar in terms of other measured variables. This approach is sometimes used in observational studies to reduce or eliminate the effects of confounding factors.

Paired samples *t*-tests are often referred to as "dependent samples *t*-tests" (as are *t*-tests on overlapping samples).

Overlapping samples

An overlapping samples *t*-test is used when there are paired samples with data missing in one or the other samples (e.g., due to selection of "Don't know" options in questionnaires or because respondents are randomly assigned to a subset question). These tests are widely used in commercial survey research (e.g., by polling companies) and are available in many standard crosstab software packages.

Calculations

Explicit expressions that can be used to carry out various *t*-tests are given below. In each case, the formula for a test statistic that either exactly follows or closely approximates a *t*-distribution under the null hypothesis is given. Also, the appropriate degrees of freedom are given in each case. Each of these statistics can be used to carry out either a one-tailed test or a two-tailed test.

Once a *t* value is determined, a p-value can be found using a table of values from Student's *t*-distribution. If the calculated p-value is below the threshold chosen for statistical significance (usually the 0.10, the 0.05, or 0.01 level), then the null hypothesis is rejected in favor of the alternative hypothesis.

One-sample *t*-test

In testing the null hypothesis that the population mean is equal to a specified value μ_0 , one uses the statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where \bar{x} is the sample mean, *s* is the sample standard deviation of the sample and *n* is the sample size. The degrees of freedom used in this test are *n* - 1. Although the parent population does not need to be normally distributed, the distribution of the population of sample means, \bar{x} , is assumed to be normal. By the central limit theorem, if the sampling of the parent population is random then the sample means will be approximately normal.^[4] (The degree of approximation will depend on how close the parent population is to a normal distribution and the sample size, *n*.)

Slope of a regression line

Suppose one is fitting the model

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

where $x_i, i = 1, \dots, n$ are known, α and β are unknown, and ε_i are independent identically normally distributed random errors with expected value 0 and unknown variance σ^2 , and $Y_i, i = 1, \dots, n$ are observed. It is desired to test the null hypothesis that the slope β is equal to some specified value β_0 (often taken to be 0, in which case the hypothesis is that *x* and *y* are unrelated).

Let

$$\hat{\alpha}, \hat{\beta} = \text{least-squares estimators,}$$

$$SE_{\hat{\alpha}}, SE_{\hat{\beta}} = \text{the standard errors of least-squares estimators.}$$

Then

$$t_{\text{score}} = \frac{\hat{\beta} - \beta_0}{SE_{\hat{\beta}}} \sim T_{n-2}$$

has a *t*-distribution with *n* - 2 degrees of freedom if the null hypothesis is true. The standard error of the slope coefficient:

$$SE_{\hat{\beta}} = \frac{\sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

can be written in terms of the residuals. Let

$$\hat{\varepsilon}_i = Y_i - \hat{y}_i = Y_i - (\hat{\alpha} + \hat{\beta}x_i) = \text{residuals} = \text{estimated errors},$$

$$\text{SSR} = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \text{sum of squares of residuals}.$$

Then t_{score} is given by:

$$t_{\text{score}} = \frac{(\hat{\beta} - \beta_0)\sqrt{n-2}}{\sqrt{\text{SSR} / \sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Independent two-sample t -test

Equal sample sizes, equal variance

This test is only used when both:

- the two sample sizes (that is, the number, n , of participants of each group) are equal;
- it can be assumed that the two distributions have the same variance.

Violations of these assumptions are discussed below.

The t statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

where

$$s_{X_1X_2} = \sqrt{\frac{1}{2}(s_{X_1}^2 + s_{X_2}^2)}$$

Here $s_{X_1X_2}$ is the grand standard deviation (or pooled standard deviation), 1 = group one, 2 = group two. $s_{X_1}^2$ and $s_{X_2}^2$ are the unbiased estimators of the variances of the two samples. The denominator of t is the standard error of the difference between two means.

For significance testing, the degrees of freedom for this test is $2n - 2$ where n is the number of participants in each group.

Unequal sample sizes, equal variance

This test is used only when it can be assumed that the two distributions have the same variance. (When this assumption is violated, see below.) The t statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_{X_1X_2} = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}.$$

Note that the formulae above are generalizations of the case where both samples have equal sizes (substitute n for n_1 and n_2).

$s_{X_1X_2}$ is an estimator of the common standard deviation of the two samples: it is defined in this way so that its square is an unbiased estimator of the common variance whether or not the population means are the same. In these formulae, n = number of participants, 1 = group one, 2 = group two. $n - 1$ is the number of degrees of freedom for either group, and the total sample size minus two (that is, $n_1 + n_2 - 2$) is the total number of degrees of freedom, which is used in significance testing.

Equal or Unequal sample sizes, unequal variances

This test, also known as Welch's *t*-test, is used only when the two population variances are not assumed to be equal (the two sample sizes may or may not be equal) and hence must be estimated separately. The *t* statistic to test whether the population means are different is calculated as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

where

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here s^2 is the unbiased estimator of the variance of the two samples, n_i = number of participants in group i , $i=1$ or 2 . Note that in this case $s_{\bar{X}_1 - \bar{X}_2}^2$ is not a pooled variance. For use in significance testing, the distribution of the test statistic is approximated as an ordinary Student's *t* distribution with the degrees of freedom calculated using

$$\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

This is known as the Welch–Satterthwaite equation. The true distribution of the test statistic actually depends (slightly) on the two unknown population variances (see Behrens–Fisher problem).

Dependent *t*-test for paired samples

This test is used when the samples are dependent; that is, when there is only one sample that has been tested twice (repeated measures) or when there are two samples that have been matched or "paired". This is an example of a paired difference test.

$$t = \frac{\bar{X}_D - \mu_0}{s_D/\sqrt{n}}$$

For this equation, the differences between all pairs must be calculated. The pairs are either one person's pre-test and post-test scores or between pairs of persons matched into meaningful groups (for instance drawn from the same family or age group: see table). The average (\bar{X}_D) and standard deviation (s_D) of those differences are used in the equation. The constant μ_0 is non-zero if you want to test whether the average of the difference is significantly different from μ_0 . The degree of freedom used is $n - 1$.

<i>Example of repeated measures</i>			
Number	Name	Test 1	Test 2
1	Mike	35%	67%
2	Melanie	50%	46%
3	Melissa	90%	86%
4	Mitchell	78%	91%

<i>Example of matched pairs</i>			
Pair	Name	Age	Test
1	John	35	250
1	Jane	36	340
2	Jimmy	22	460
2	Jessy	21	200

Worked examples

Let A_1 denote a set obtained by taking 6 random samples out of a larger set:

$$A_1 = \{30.02, 29.99, 30.11, 29.97, 30.01, 29.99\}$$

and let A_2 denote a second set obtained similarly:

$$A_2 = \{29.89, 29.93, 29.72, 29.98, 30.02, 29.98\}$$

These could be, for example, the weights of screws that were chosen out of a bucket.

We will carry out tests of the null hypothesis that the means of the populations from which the two samples were taken are equal.

The difference between the two sample means, each denoted by \bar{X}_i , which appears in the numerator for all the two-sample testing approaches discussed above, is

$$\bar{X}_1 - \bar{X}_2 = 0.095.$$

The sample standard deviations for the two samples are approximately 0.05 and 0.11, respectively. For such small samples, a test of equality between the two population variances would not be very powerful. Since the sample sizes are equal, the two forms of the two sample t -test will perform similarly in this example.

Unequal variances

If the approach for unequal variances (discussed above) is followed, the results are

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \approx 0.0485$$

and

$$df \approx 7.03.$$

The test statistic is approximately 1.959. The two-tailed test p-value is approximately 0.091 and the one-tailed p-value is approximately 0.045.

Equal variances

If the approach for equal variances (discussed above) is followed, the results are

$$S_{X_1, X_2} \approx 0.084$$

and

$$df = 10.$$

Since the sample sizes are equal (both are 6), the test statistic is again approximately equal to 1.959. Since the degrees of freedom is different from what it is in the unequal variances test, the p-values will differ slightly from what was found above. Here, the two-tailed p-value is approximately 0.078, and the one-tailed p-value is approximately 0.039. Thus if there is good reason to believe that the population variances are equal, the results become somewhat more suggestive of a difference in the mean weights for the two populations of screws.

Alternatives to the *t*-test for location problems

The *t*-test provides an exact test for the equality of the means of two normal populations with unknown, but equal, variances. (The Welch's *t*-test is a nearly exact test for the case where the data are normal but the variances may differ.) For moderately large samples and a one tailed test, the *t* is relatively robust to moderate violations of the normality assumption.

For exactness, the *t*-test and *Z*-test require normality of the sample means, and the *t*-test additionally requires that the sample variance follows a scaled χ^2 distribution, and that the sample mean and sample variance be statistically independent. Normality of the individual data values is not required if these conditions are met. By the central limit theorem, sample means of moderately large samples are often well-approximated by a normal distribution even if the data are not normally distributed. For non-normal data, the distribution of the sample variance may deviate substantially from a χ^2 distribution. However, if the sample size is large, Slutsky's theorem implies that the distribution of the sample variance has little effect on the distribution of the test statistic. If the data are substantially non-normal and the sample size is small, the *t*-test can give misleading results. See Location test for Gaussian scale mixture distributions for some theory related to one particular family of non-normal distributions.

When the normality assumption does not hold, a non-parametric alternative to the *t*-test can often have better statistical power. For example, for two independent samples when the data distributions are asymmetric (that is, the distributions are skewed) or the distributions have large tails, then the Wilcoxon rank-sum test (also known as the Mann–Whitney U test) can have three to four times higher power than the *t*-test. The nonparametric counterpart to the paired samples *t* test is the Wilcoxon signed-rank test for paired samples. For a discussion on choosing between the *t* and nonparametric alternatives, see Sawilowsky.

One-way analysis of variance generalizes the two-sample *t*-test when the data belong to more than two groups.

Multivariate testing

A generalization of Student's *t* statistic, called Hotelling's *T*-square statistic, allows for the testing of hypotheses on multiple (often correlated) measures within the same sample. For instance, a researcher might submit a number of subjects to a personality test consisting of multiple personality scales (e.g. the Minnesota Multiphasic Personality Inventory). Because measures of this type are usually positively correlated, it is not advisable to conduct separate univariate *t*-tests to test hypotheses, as these would neglect the covariance among measures and inflate the chance of falsely rejecting at least one hypothesis (Type I error). In this case a single multivariate test is preferable for hypothesis testing. Fisher's Method for combining multiple tests with *alpha* reduced for positive correlation among tests is one. Another is Hotelling's T^2 statistic follows a T^2 distribution. However, in practice the distribution is rarely used, since tabulated values for T^2 are hard to find. Usually, T^2 is converted instead to an *F* statistic.

One-sample T^2 test

For a one-sample multivariate test, the hypothesis is that the mean vector (μ) is equal to a given vector (μ_0). The test statistic is Hotelling's T^2 :

$$T^2 = n(\bar{\mathbf{x}} - \mu_0)' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \mu_0)$$

where n is the sample size, $\bar{\mathbf{x}}$ is the vector of column means and \mathbf{S} is a $m \times m$ sample covariance matrix.

Two-sample T^2 test

For a two-sample multivariate test, the hypothesis is that the mean vectors (μ_1, μ_2) of two samples are equal. The test statistic is Hotelling's 2-sample T^2 :

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Software implementations

Many spreadsheet programs and statistics packages, such as QtiPlot, OpenOffice.org Calc, LibreOffice Calc, Microsoft Excel, SAS, SPSS, Stata, DAP, gretl, R, Python ([5]), PSPP, and Minitab, include implementations of Student's *t*-test.

Language/Program	Function	Notes
Microsoft Excel pre 2010	<code>TTEST(array1, array2, tails, type)</code>	See [6]
Microsoft Excel 2010 and later	<code>T.TEST(array1, array2, tails, type)</code>	See [7]
OpenOffice.org	<code>TTEST(data1; data2; mode; type)</code>	
Python	<code>scipy.stats.ttest_ind(a, b, axis=0, equal_var=True)</code>	See [5]
R	<code>t.test(data1, data2)</code>	
SAS	PROC TTEST	See [8]

Notes

- [1] Richard Mankiewicz, *The Story of Mathematics* (Princeton University Press), p.158.
- [2] http://www.aliquote.org/cours/2012_biomed/biblio/Student1908.pdf
- [3] John A. Rice (2006), *Mathematical Statistics and Data Analysis*, Third Edition, Duxbury Advanced.
- [4] George Box, William Hunter, and J. Stuart Hunter, "Statistics for Experimenters", ISBN 978-0471093152, pp. 66-67.
- [5] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html
- [6] <http://office.microsoft.com/en-us/excel-help/ttest-HP005209325.aspx>
- [7] <http://office.microsoft.com/en-us/excel-help/t-test-function-HA102753135.aspx>
- [8] <http://www.sas.com/offices/europe/belux/pdf/academic/ttest.pdf>

References

- O'Mahony, Michael (1986). *Sensory Evaluation of Food: Statistical Methods and Procedures*. CRC Press. p. 487. ISBN 0-8247-7337-3.
- Press, William H.; Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing* (<http://www.nr.com/>). Cambridge University Press. pp. p. 616 (<http://www.nrbook.com/a/bookcpdf/c14-2.pdf>). ISBN 0-521-43108-5.

Further reading

- Boneau, C. Alan (1960). "The effects of violations of assumptions underlying the *t* test". *Psychological Bulletin* **57** (1): 49–64. doi: 10.1037/h0041412 (<http://dx.doi.org/10.1037/h0041412>)
- Edgell, Stephen E., & Noon, Sheila M (1984). "Effect of violation of normality on the *t* test of the correlation coefficient". *Psychological Bulletin* **95** (3): 576–583. doi: 10.1037/0033-2909.95.3.576 (<http://dx.doi.org/10.1037/0033-2909.95.3.576>).

External links

- Hazewinkel, Michiel, ed. (2001), "Student test" (<http://www.encyclopediaofmath.org/index.php?title=p/s090720>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- A conceptual article on the Student's *t*-test (http://www.socialresearchmethods.net/kb/stat_t.php)

Contingency table

In statistics, a **contingency table** (also referred to as **cross tabulation** or **cross tab**) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. The term *contingency table* was first used by Karl Pearson in "On the Theory of Contingency and Its Relation to Association and Normal Correlation",^[1] part of the *Drapers' Company Research Memoirs Biometric Series I* published in 1904.

A crucial problem of multivariate statistics is finding (direct-)dependence structure underlying the variables contained in high dimensional contingency tables. If some of the conditional independences are revealed, then even the storage of the data can be done in a smarter way (see Lauritzen (2002)). In order to do this one can use information theory concepts, which gain the information only from the distribution of probability, which can be expressed easily from the contingency table by the relative frequencies.

Example

Suppose that we have two variables, sex (male or female) and handedness (right- or left-handed). Further suppose that 100 individuals are randomly sampled from a very large population as part of a study of sex differences in handedness. A contingency table can be created to display the numbers of individuals who are male and right-handed, male and left-handed, female and right-handed, and female and left-handed. Such a contingency table is shown below.

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Totals	87	13	100

The numbers of the males, females, and right- and left-handed individuals are called **marginal totals**. The **grand total**, i.e., the total number of individuals represented in the contingency table, is the number in the bottom right corner.

The table allows us to see at a glance that the proportion of men who are right-handed is about the same as the proportion of women who are right-handed although the proportions are not identical. The significance of the difference between the two proportions can be assessed with a variety of statistical tests including Pearson's chi-squared test, the *G*-test, Fisher's exact test, and Barnard's test, provided the entries in the table represent individuals randomly sampled from the population about which we want to draw a conclusion. If the proportions of individuals in the different columns vary significantly between rows (or vice versa), we say that there is a *contingency* between the two variables. In other words, the two variables are *not* independent. If there is no contingency, we say that the two variables are *independent*.

The example above is the simplest kind of contingency table, a table in which each variable has only two levels; this is called a 2 x 2 contingency table. In principle, any number of rows and columns may be used. There may also be more than two variables, but higher order contingency tables are difficult to represent on paper. The relation between ordinal variables, or between ordinal and categorical variables, may also be represented in contingency tables, although such a practice is rare.

Measures of association

The degree of association between the two variables can be assessed by a number of coefficients: the simplest is the phi coefficient defined by

$$\phi = \sqrt{\frac{\chi^2}{N}},$$

where χ^2 is derived from Pearson's chi-squared test, and N is the grand total of observations. ϕ varies from 0 (corresponding to no association between the variables) to 1 or -1 (complete association or complete inverse association). This coefficient can only be calculated for frequency data represented in 2 x 2 tables. ϕ can reach a minimum value -1.00 and a maximum value of 1.00 *only* when every marginal proportion is equal to .50 (and two diagonal cells are empty). Otherwise, the phi coefficient cannot reach those minimal and maximal values.^[2]

Alternatives include the tetrachoric correlation coefficient (also only applicable to 2 x 2 tables), the *contingency coefficient C*, and **Cramér's V**.

C suffers from the disadvantage that it does not reach a maximum of 1 or the minimum of -1; the highest it can reach in a 2 x 2 table is .707; the maximum it can reach in a 4 x 4 table is 0.870. It can reach values closer to 1 in contingency tables with more categories. It should, therefore, not be used to compare associations among tables with different numbers of categories.^[3] Moreover, it does not apply to asymmetrical tables (those where the numbers of row and columns are not equal).

The formulae for the C and V coefficients are:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \text{ and}$$

$$V = \sqrt{\frac{\chi^2}{N(k - 1)}},$$

k being the number of rows or the number of columns, whichever is less.

C can be adjusted so it reaches a maximum of 1 when there is complete association in a table of any number of rows and columns by dividing C by $\sqrt{\frac{k - 1}{k}}$ (recall that C only applies to tables in which the number of rows is equal to the number of columns and therefore equal to k).

The tetrachoric correlation coefficient assumes that the variable underlying each dichotomous measure is normally distributed.^[4] The tetrachoric correlation coefficient provides "a convenient measure of [the Pearson product-moment] correlation when graduated measurements have been reduced to two categories."^[5] The tetrachoric correlation should not be confused with the Pearson product-moment correlation coefficient computed by assigning, say, values 0 and 1 to represent the two levels of each variable (which is mathematically equivalent to the phi coefficient). An extension of the tetrachoric correlation to tables involving variables with more than two levels is the polychoric correlation coefficient.

The **Lambda coefficient** is a measure of the strength of association of the cross tabulations when the variables are measured at the nominal level. Values range from 0 (no association) to 1 (the theoretical maximum possible association). **Asymmetric lambda** measures the percentage improvement in predicting the dependent variable. **Symmetric lambda** measures the percentage improvement when prediction is done in both directions.

The uncertainty coefficient is another measure for variables at the nominal level.

The values range from -1 (100% negative association, or perfect inversion) to +1 (100% positive association, or perfect agreement). A value of zero indicates the absence of association.

- **Gamma test:** No adjustment for either table size or ties.
- **Kendall tau:** Adjustment for ties.

- **Tau b:** For square tables.
- **Tau c:** For rectangular tables.

References

- [1] <http://ia600408.us.archive.org/18/items/cu31924003064833/cu31924003064833.pdf>
- [2] Ferguson, G. A. (1966). *Statistical analysis in psychology and education*. New York: McGraw-Hill.
- [3] Smith, S. C., & Albaum, G. S. (2004) *Fundamentals of marketing research*. Sage: Thousand Oaks, CA. p. 631
- [4] Ferguson.
- [5] Ferguson, p. 244
- Andersen, Erling B. 1980. *Discrete Statistical Models with Social Science Applications*. North Holland, 1980.
 - Bishop, Y. M. M.; Fienberg, S. E.; Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press. ISBN 978-0-262-02113-5. MR 381130 (<http://www.ams.org/mathscinet-getitem?mr=381130>).
 - Christensen, Ronald (1997). *Log-linear models and logistic regression*. Springer Texts in Statistics (Second ed.). New York: Springer-Verlag. pp. xvi+483. ISBN 0-387-98247-7. MR 1633357 (<http://www.ams.org/mathscinet-getitem?mr=1633357>).
 - Lauritzen, Steffen L. (2002 electronic (1979, 1982, 1989)). *Lectures on Contingency Tables* (<http://www.stats.ox.ac.uk/~steffen/papers/cont.pdf>) (updated electronic version of the (University of Aalborg) 3rd (1989) ed.).
 - Gokhale, D. V.; Kullback, Solomon (1978). *The Information in Contingency Tables*. Marcel Dekker. ISBN 0-824-76698-9.

External links

- On-line analysis of contingency tables: calculator with examples (<http://www.physics.csbsju.edu/stats/contingency.html>)
- Interactive cross tabulation, chi-squared independent test & tutorial (<http://people.revoledu.com/kardi/tutorial/Questionnaire/ContingencyTable.html>)
- Fisher and chi-squared calculator of 2×2 contingency table (<http://statpages.org/ctab2x2.html>)
- More Correlation Coefficients (<http://www.andrews.edu/~calkins/math/edrm611/edrm13.htm>)
- Nominal Association: Phi, Contingency Coefficient, Tschuprow's T, Cramer's V, Lambda, Uncertainty Coefficient (<http://www2.chass.ncsu.edu/garson/pa765/assocnominal.htm>)
- Customer Insight com Cross Tabulation (<http://www.custominsight.com/articles/crosstab-sample.asp>)
- The POWERMUTT Project: IV. DISPLAYING CATEGORICAL DATA (http://www.csupomona.edu/~jlkorey/POWERMUTT/Topics/displaying_categorical_data.html)
- StATS: Steves Attempt to Teach Statistics Odds ratio versus relative risk (January 9, 2001) (<http://www.childrensmemory.org/stats/journal/oddsratio.asp>)
- Epi Info Community Health Assessment Tutorial Lesson 5 Analysis: Creating Statistics (ftp://ftp.cdc.gov/pub/Software/epi_info/EIHAT_WEB/Lesson5AnalysisCreatingStatistics.pdf)

Analysis of variance

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences between group means and their associated procedures (such as "variation" among and between groups). In ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the *t*-test to more than two groups. Doing multiple two-sample *t*-tests would result in an increased chance of committing a type I error. For this reason, ANOVAs are useful in comparing (testing) three or more means (groups or variables) for statistical significance.

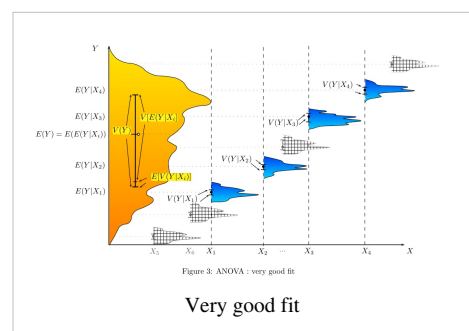
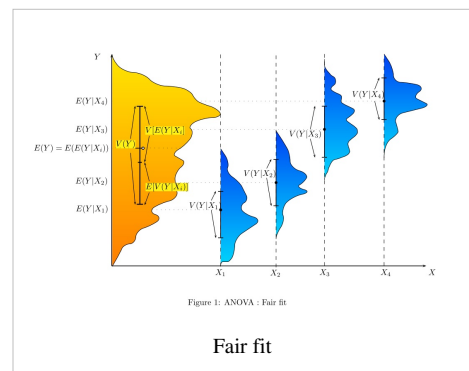
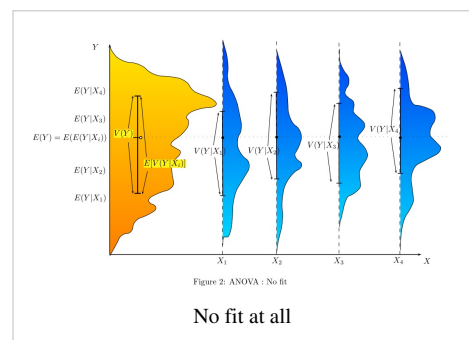
Motivating example

The analysis of variance can be used as an exploratory tool to explain observations. A dog show provides an example. A dog show is not a random sampling of the breed. It is typically limited to dogs that are male, adult, pure-bred and exemplary. A histogram of dog weights from a show might plausibly be rather complex, like the yellow-orange distribution shown in the illustrations. An attempt to explain the weight distribution by dividing the dog population into groups (young vs old)(short-haired vs long-haired) would probably be a failure (no fit at all). The groups (shown in blue) have a large variance and the means are very close. An attempt to explain the weight distribution by (pet vs working breed)(less athletic vs more athletic) would probably be somewhat more successful (fair fit). The heaviest show dogs are likely to be big strong working breeds. An attempt to explain weight by breed is likely to produce a very good fit. All Chihuahuas are light and all St Bernards are heavy. The difference in weights between Setters and Pointers does not justify separate breeds. The analysis of variance provides the formal tools to justify these intuitive judgments. A common use of the method is the analysis of experimental data or the development of models. The method has some advantages over correlation: not all of the data must be numeric and one result of the method is a judgment in the confidence in an explanatory relationship.

Background and terminology

ANOVA is a particular form of statistical hypothesis testing heavily used in the analysis of experimental data. A statistical hypothesis test is a method of making decisions using data. A test result (calculated from the null hypothesis and the sample) is called statistically significant if it is deemed unlikely to have occurred by chance, *assuming the truth of the null hypothesis*. A statistically significant result (when a probability (p-value) is less than a threshold (significance level)) justifies the rejection of the null hypothesis, but only if the a priori probability of the null hypothesis is not high.

In the typical application of ANOVA, the null hypothesis is that all groups are simply random samples of the same population. This implies that all treatments have the same effect (perhaps none). Rejecting the null hypothesis implies that different treatments result in altered effects.



By construction, hypothesis testing limits the rate of Type I errors (false positives leading to false scientific claims) to a significance level. Experimenters also wish to limit Type II errors (false negatives resulting in missed scientific discoveries). The Type II error rate is a function of several things including sample size (positively correlated with experiment cost), significance level (when the standard of proof is high, the chances of overlooking a discovery are also high) and effect size (when the effect is obvious to the casual observer, Type II error rates are low).

The terminology of ANOVA is largely from the statistical design of experiments. The experimenter adjusts factors and measures responses in an attempt to determine an effect. Factors are assigned to experimental units by a combination of randomization and blocking to ensure the validity of the results. Blinding keeps the weighing impartial. Responses show a variability that is partially the result of the effect and is partially random error.

ANOVA is the synthesis of several ideas and it is used for multiple purposes. As a consequence, it is difficult to define concisely or precisely.

"Classical ANOVA for balanced data does three things at once:

1. As exploratory data analysis, an ANOVA is an organization of an additive data decomposition, and its sums of squares indicate the variance of each component of the decomposition (or, equivalently, each set of terms of a linear model).
2. Comparisons of mean squares, along with F-tests ... allow testing of a nested sequence of models.
3. Closely related to the ANOVA is a linear model fit with coefficient estimates and standard errors."^[1]

In short, ANOVA is a statistical tool used in several ways to develop and confirm an explanation for the observed data.

Additionally:

- It is computationally elegant and relatively robust against violations of its assumptions.
2. ANOVA provides industrial strength (multiple sample comparison) statistical analysis.
 3. It has been adapted to the analysis of a variety of experimental designs.

As a result: ANOVA "has long enjoyed the status of being the **most used** (some would say abused) statistical technique in psychological research."^[2] ANOVA "is probably the **most useful** technique in the field of statistical inference."^[3]

ANOVA is difficult to teach, particularly for complex experiments, with split-plot designs being notorious.^[4] In some cases the proper application of the method is best determined by problem pattern recognition followed by the consultation of a classic authoritative test.^[5]

Design-of-experiments terms

(Condensed from the NIST Engineering Statistics handbook: Section 5.7. A Glossary of DOE Terminology.)

Balanced design

An experimental design where all cells (i.e. treatment combinations) have the same number of observations.

Blocking

A schedule for conducting treatment combinations in an experimental study such that any effects on the experimental results due to a known change in raw materials, operators, machines, etc., become concentrated in the levels of the blocking variable. The reason for blocking is to isolate a systematic effect and prevent it from obscuring the main effects. Blocking is achieved by restricting randomization.

Design

A set of experimental runs which allows the fit of a particular model and the estimate of effects.

DOE

Design of experiments. An approach to problem solving involving collection of data that will support valid, defensible, and supportable conclusions.

Effect

How changing the settings of a factor changes the response. The effect of a single factor is also called a main effect.

Error

Unexplained variation in a collection of observations. DOE's typically require understanding of both random error and lack of fit error.

Experimental unit

The entity to which a specific treatment combination is applied.

Factors

Process inputs an investigator manipulates to cause a change in the output.

Lack-of-fit error

Error that occurs when the analysis omits one or more important terms or factors from the process model. Including replication in a DOE allows separation of experimental error into its components: lack of fit and random (pure) error.

Model

Mathematical relationship which relates changes in a given response to changes in one or more factors.

Random error

Error that occurs due to natural variation in the process. Random error is typically assumed to be normally distributed with zero mean and a constant variance. Random error is also called experimental error.

Randomization

A schedule for allocating treatment material and for conducting treatment combinations in a DOE such that the conditions in one run neither depend on the conditions of the previous run nor predict the conditions in the subsequent runs.^[6]

Replication

Performing the same treatment combination more than once. Including replication allows an estimate of the random error independent of any lack of fit error.

Responses

The output(s) of a process. Sometimes called dependent variable(s).

Treatment

A treatment is a specific combination of factor levels whose effect is to be compared with other treatments.

Classes of models

There are three classes of models used in the analysis of variance, and these are outlined here.

Fixed-effects models

The fixed-effects model of analysis of variance applies to situations in which the experimenter applies one or more treatments to the subjects of the experiment to see if the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole.

Random-effects models

Random effects models are used when the treatments are not fixed. This occurs when the various factor levels are sampled from a larger population. Because the levels themselves are random variables, some assumptions and the method of contrasting the treatments (a multi-variable generalization of simple differences) differ from the fixed-effects model.^[7]

Mixed-effects models

A mixed-effects model contains experimental factors of both fixed and random-effects types, with appropriately different interpretations and analysis for the two types.

Example: Teaching experiments could be performed by a university department to find a good introductory textbook, with each text considered a treatment. The fixed-effects model would compare a list of candidate texts. The random-effects model would determine whether important differences exist among a list of randomly selected texts. The mixed-effects model would compare the (fixed) incumbent texts to randomly selected alternatives.

Defining fixed and random effects has proven elusive, with competing definitions arguably leading toward a linguistic quagmire.^[8]

Assumptions of ANOVA

The analysis of variance has been studied from several approaches, the most common of which uses a linear model that relates the response to the treatments and blocks. Note that the model is linear in parameters but may be nonlinear across factor levels. Interpretation is easy when data is balanced across factors but much deeper understanding is needed for unbalanced data.

Textbook analysis using a normal distribution

The analysis of variance can be presented in terms of a linear model, which makes the following assumptions about the probability distribution of the responses:^{[9][10]}

- Independence of observations – this is an assumption of the model that simplifies the statistical analysis.
- Normality – the distributions of the residuals are normal.
- Equality (or "homogeneity") of variances, called homoscedasticity — the variance of data in groups should be the same.

The separate assumptions of the textbook model imply that the errors are independently, identically, and normally distributed for fixed effects models, that is, that the errors (ε 's) are independent and

$$\varepsilon \sim N(0, \sigma^2).$$

Randomization-based analysis

In a randomized controlled experiment, the treatments are randomly assigned to experimental units, following the experimental protocol. This randomization is objective and declared before the experiment is carried out. The objective random-assignment is used to test the significance of the null hypothesis, following the ideas of C. S. Peirce and Ronald A. Fisher. This design-based analysis was discussed and developed by Francis J. Anscombe at Rothamsted Experimental Station and by Oscar Kempthorne at Iowa State University.^[11] Kempthorne and his students make an assumption of *unit treatment additivity*, which is discussed in the books of Kempthorne and David R. Cox.^[citation needed]

Unit-treatment additivity

In its simplest form, the assumption of unit-treatment additivity^[12] states that the observed response $y_{i,j}$ from experimental unit i when receiving treatment j can be written as the sum of the unit's response y_i and the treatment-effect t_j , that is^{[13][14][15]}

$$y_{i,j} = y_i + t_j.$$

The assumption of unit-treatment additivity implies that, for every treatment j , the j th treatment have exactly the same effect t_j on every experiment unit.

The assumption of unit treatment additivity usually cannot be directly falsified, according to Cox and Kempthorne. However, many *consequences* of treatment-unit additivity can be falsified. For a randomized experiment, the assumption of unit-treatment additivity *implies* that the variance is constant for all treatments. Therefore, by contraposition, a necessary condition for unit-treatment additivity is that the variance is constant.

The use of unit treatment additivity and randomization is similar to the design-based inference that is standard in finite-population survey sampling.

Derived linear model

Kempthorne uses the randomization-distribution and the assumption of *unit treatment additivity* to produce a *derived linear model*, very similar to the textbook model discussed previously.^[16] The test statistics of this derived linear model are closely approximated by the test statistics of an appropriate normal linear model, according to approximation theorems and simulation studies.^[17] However, there are differences. For example, the randomization-based analysis results in a small but (strictly) negative correlation between the observations.^{[18][19]} In the randomization-based analysis, there is *no assumption* of a *normal* distribution and certainly *no assumption* of *independence*. On the contrary, *the observations are dependent!*

The randomization-based analysis has the disadvantage that its exposition involves tedious algebra and extensive time. Since the randomization-based analysis is complicated and is closely approximated by the approach using a normal linear model, most teachers emphasize the normal linear model approach. Few statisticians object to model-based analysis of balanced randomized experiments.

Statistical models for observational data

However, when applied to data from non-randomized experiments or observational studies, model-based analysis lacks the warrant of randomization.^[20] For observational data, the derivation of confidence intervals must use *subjective* models, as emphasized by Ronald A. Fisher and his followers. In practice, the estimates of treatment-effects from observational studies generally are often inconsistent. In practice, "statistical models" and observational data are useful for suggesting hypotheses that should be treated very cautiously by the public.^[21]

Summary of assumptions

The normal-model based ANOVA analysis assumes the independence, normality and homogeneity of the variances of the residuals. The randomization-based analysis assumes only the homogeneity of the variances of the residuals (as a consequence of unit-treatment additivity) and uses the randomization procedure of the experiment. Both these analyses require homoscedasticity, as an assumption for the normal-model analysis and as a consequence of randomization and additivity for the randomization-based analysis.

However, studies of processes that change variances rather than means (called dispersion effects) have been successfully conducted using ANOVA.^[22] There are *no* necessary assumptions for ANOVA in its full generality, but the F-test used for ANOVA hypothesis testing has assumptions and practical limitations which are of continuing interest.

Problems which do not satisfy the assumptions of ANOVA can often be transformed to satisfy the assumptions. The property of unit-treatment additivity is not invariant under a "change of scale", so statisticians often use transformations to achieve unit-treatment additivity. If the response variable is expected to follow a parametric family of probability distributions, then the statistician may specify (in the protocol for the experiment or observational study) that the responses be transformed to stabilize the variance.^[23] Also, a statistician may specify that logarithmic transforms be applied to the responses, which are believed to follow a multiplicative model.^[24] According to Cauchy's functional equation theorem, the logarithm is the only continuous transformation that transforms real multiplication to addition^[citation needed].

Characteristics of ANOVA

ANOVA is used in the analysis of comparative experiments, those in which only the difference in outcomes is of interest. The statistical significance of the experiment is determined by a ratio of two variances. This ratio is independent of several possible alterations to the experimental observations: Adding a constant to all observations does not alter significance. Multiplying all observations by a constant does not alter significance. So ANOVA statistical significance results are independent of constant bias and scaling errors as well as the units used in expressing observations. In the era of mechanical calculation it was common to subtract a constant from all observations (when equivalent to dropping leading digits) to simplify data entry.^{[25][26]} This is an example of data coding.

Logic of ANOVA

The calculations of ANOVA can be characterized as computing a number of means and variances, dividing two variances and comparing the ratio to a handbook value to determine statistical significance. Calculating a treatment effect is then trivial, "the effect of any treatment is estimated by taking the difference between the mean of the observations which receive the treatment and the general mean."^[27]

Partitioning of the sum of squares

ANOVA uses traditional standardized terminology. The definitional equation of sample variance is $s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$, where the divisor is called the degrees of freedom (DF), the summation is called the sum of squares (SS), the result is called the mean square (MS) and the squared terms are deviations from the sample mean. ANOVA estimates 3 sample variances: a total variance based on all the observation deviations from the grand mean, an error variance based on all the observation deviations from their appropriate treatment means and a treatment variance. The treatment variance is based on the deviations of treatment means from the grand mean, the result being multiplied by the number of observations in each treatment to account for the difference between the variance of observations and the variance of means.

The fundamental technique is a partitioning of the total sum of squares SS into components related to the effects used in the model. For example, the model for a simplified ANOVA with one type of treatment at different levels.

$$SS_{\text{Total}} = SS_{\text{Error}} + SS_{\text{Treatments}}$$

The number of degrees of freedom DF can be partitioned in a similar way: one of these components (that for error) specifies a chi-squared distribution which describes the associated sum of squares, while the same is true for "treatments" if there is no treatment effect.

$$DF_{\text{Total}} = DF_{\text{Error}} + DF_{\text{Treatments}}$$

See also Lack-of-fit sum of squares.

The F-test

The F-test is used for comparing the factors of the total deviation. For example, in one-way, or single-factor ANOVA, statistical significance is tested for by comparing the F test statistic

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

$$F = \frac{MS_{\text{Treatments}}}{MS_{\text{Error}}} = \frac{SS_{\text{Treatments}}/(I - 1)}{SS_{\text{Error}}/(n_T - I)}$$

where MS is mean square, I = number of treatments and n_T = total number of cases

to the F-distribution with $I - 1$, $n_T - I$ degrees of freedom. Using the F-distribution is a natural candidate because the test statistic is the ratio of two scaled sums of squares each of which follows a scaled chi-squared distribution.

The expected value of F is $1 + n\sigma_{\text{Treatment}}^2/\sigma_{\text{Error}}^2$ (where n is the treatment sample size) which is 1 for no treatment effect. As values of F increase above 1 the evidence is increasingly inconsistent with the null hypothesis. Two apparent experimental methods of increasing F are increasing the sample size and reducing the error variance by tight experimental controls.

The textbook method of concluding the hypothesis test is to compare the observed value of F with the critical value of F determined from tables. The critical value of F is a function of the numerator degrees of freedom, the denominator degrees of freedom and the significance level (α). If $F \geq F_{\text{Critical}}(\text{Numerator DF, Denominator DF, } \alpha)$ then reject the null hypothesis.

The computer method calculates the probability (p-value) of a value of F greater than or equal to the observed value. The null hypothesis is rejected if this probability is less than or equal to the significance level (α). The two methods produce the same result.

The ANOVA F-test is known to be nearly optimal in the sense of minimizing false negative errors for a fixed rate of false positive errors (maximizing power for a fixed significance level). To test the hypothesis that all treatments have exactly the same effect, the F-test's p-values closely approximate the permutation test's p-values: The approximation is particularly close when the design is balanced.^[28] Such permutation tests characterize tests with maximum power against all alternative hypotheses, as observed by Rosenbaum.^[29] The ANOVA F-test (of the null-hypothesis that all treatments have exactly the same effect) is recommended as a practical test, because of its robustness against many alternative distributions.^{[30][31]}

Extended logic

ANOVA consists of separable parts; partitioning sources of variance and hypothesis testing can be used individually. ANOVA is used to support other statistical tools. Regression is first used to fit more complex models to data, then ANOVA is used to compare models with the objective of selecting simple(r) models that adequately describe the data. "Such models could be fit without any reference to ANOVA, but ANOVA tools could then be used to make some sense of the fitted models, and to test hypotheses about batches of coefficients."^[32] "[W]e think of the analysis of variance as a way of understanding and structuring multilevel models—not as an alternative to regression but as a tool for summarizing complex high-dimensional inferences ..."

ANOVA for a single factor

The simplest experiment suitable for ANOVA analysis is the completely randomized experiment with a single factor. More complex experiments with a single factor involve constraints on randomization and include completely randomized blocks and Latin squares (and variants: Graeco-Latin squares, etc.). The more complex experiments share many of the complexities of multiple factors. A relatively complete discussion of the analysis (models, data summaries, ANOVA table) of the completely randomized experiment is available.

ANOVA for multiple factors

ANOVA generalizes to the study of the effects of multiple factors. When the experiment includes observations at all combinations of levels of each factor, it is termed factorial. Factorial experiments are more efficient than a series of single factor experiments and the efficiency grows as the number of factors increases.^[33] Consequently, factorial designs are heavily used.

The use of ANOVA to study the effects of multiple factors has a complication. In a 3-way ANOVA with factors x , y and z , the ANOVA model includes terms for the main effects (x , y , z) and terms for interactions (xy , xz , yz , xyz). All terms require hypothesis tests. The proliferation of interaction terms increases the risk that some hypothesis test will produce a false positive by chance. Fortunately, experience says that high order interactions are rare.^[34] The ability to detect interactions is a major advantage of multiple factor ANOVA. Testing one factor at a time hides interactions, but produces apparently inconsistent experimental results.

Caution is advised when encountering interactions; Test interaction terms first and expand the analysis beyond ANOVA if interactions are found. Texts vary in their recommendations regarding the continuation of the ANOVA procedure after encountering an interaction. Interactions complicate the interpretation of experimental data. Neither the calculations of significance nor the estimated treatment effects can be taken at face value. "A significant interaction will often mask the significance of main effects."^[35] Graphical methods are recommended to enhance understanding. Regression is often useful. A lengthy discussion of interactions is available in Cox (1958).^[36] Some interactions can be removed (by transformations) while others cannot.

A variety of techniques are used with multiple factor ANOVA to reduce expense. One technique used in factorial designs is to minimize replication (possibly no replication with support of analytical trickery) and to combine groups when effects are found to be statistically (or practically) insignificant. An experiment with many insignificant factors may collapse into one with a few factors supported by many replications.^[37]

Worked numeric examples

Several fully worked numerical examples are available. A simple case uses one-way (a single factor) analysis. A more complex case uses two-way (two-factor) analysis.

Associated analysis

Some analysis is required in support of the *design* of the experiment while other analysis is performed after changes in the factors are formally found to produce statistically significant changes in the responses. Because experimentation is iterative, the results of one experiment alter plans for following experiments.

Preparatory analysis

The number of experimental units

In the design of an experiment, the number of experimental units is planned to satisfy the goals of the experiment. Experimentation is often sequential.

Early experiments are often designed to provide mean-unbiased estimates of treatment effects and of experimental error. Later experiments are often designed to test a hypothesis that a treatment effect has an important magnitude; in this case, the number of experimental units is chosen so that the experiment is within budget and has adequate power, among other goals.

Reporting sample size analysis is generally required in psychology. "Provide information on sample size and the process that led to sample size decisions."^[38] The analysis, which is written in the experimental protocol before the experiment is conducted, is examined in grant applications and administrative review boards.

Besides the power analysis, there are less formal methods for selecting the number of experimental units. These include graphical methods based on limiting the probability of false negative errors, graphical methods based on an expected variation increase (above the residuals) and methods based on achieving a desired confident interval.^[39]

Power analysis

Power analysis is often applied in the context of ANOVA in order to assess the probability of successfully rejecting the null hypothesis if we assume a certain ANOVA design, effect size in the population, sample size and significance level. Power analysis can assist in study design by determining what sample size would be required in order to have a reasonable chance of rejecting the null hypothesis when the alternative hypothesis is true.^{[40][41][42][43]}

Effect size

Several standardized measures of effect have been proposed for ANOVA to summarize the strength of the association between a predictor(s) and the dependent variable (e.g., η^2 , ω^2 , or f^2) or the overall standardized difference (Ψ) of the complete model. Standardized effect-size estimates facilitate comparison of findings across studies and disciplines. However, while standardized effect sizes are commonly used in much of the professional literature, a non-standardized measure of effect size that has immediately "meaningful" units may be preferable for reporting purposes.^[44]

Followup analysis

It is always appropriate to carefully consider outliers. They have a disproportionate impact on statistical conclusions and are often the result of errors.

Model confirmation

It is prudent to verify that the assumptions of ANOVA have been met. Residuals are examined or analyzed to confirm homoscedasticity and gross normality.^[45] Residuals should have the appearance of (zero mean normal distribution) noise when plotted as a function of anything including time and modeled data values. Trends hint at interactions among factors or among observations. One rule of thumb: "If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumption of equal standard deviations and our results will still be approximately correct."^[46]

Follow-up tests

A statistically significant effect in ANOVA is often followed up with one or more different follow-up tests. This can be done in order to assess which groups are different from which other groups or to test various other focused hypotheses. Follow-up tests are often distinguished in terms of whether they are planned (a priori) or post hoc. Planned tests are determined before looking at the data and post hoc tests are performed after looking at the data.

Often one of the "treatments" is none, so the treatment group can act as a control. Dunnett's test (a modification of the t-test) tests whether each of the other treatment groups has the same mean as the control.^[47]

Post hoc tests such as Tukey's range test most commonly compare every group mean with every other group mean and typically incorporate some method of controlling for Type I errors. Comparisons, which are most commonly planned, can be either simple or compound. Simple comparisons compare one group mean with one other group mean. Compound comparisons typically compare two sets of groups means where one set has two or more groups (e.g., compare average group means of group A, B and C with group D). Comparisons can also look at tests of trend, such as linear and quadratic relationships, when the independent variable involves ordered levels.

Following ANOVA with pair-wise multiple-comparison tests has been criticized on several grounds.^[48] There are many such tests (10 in one table) and recommendations regarding their use are vague or conflicting.^{[49][50]}

Study designs and ANOVAs

There are several types of ANOVA. Many statisticians base ANOVA on the design of the experiment,^[51] especially on the protocol that specifies the random assignment of treatments to subjects; the protocol's description of the assignment mechanism should include a specification of the structure of the treatments and of any blocking. It is also common to apply ANOVA to observational data using an appropriate statistical model.^[citation needed]

Some popular designs use the following types of ANOVA:

- One-way ANOVA is used to test for differences among two or more independent groups (means), e.g. different levels of urea application in a crop. Typically, however, the one-way ANOVA is used to test for differences among at least three groups, since the two-group case can be covered by a t-test. When there are only two means to compare, the t-test and the ANOVA F-test are equivalent; the relation between ANOVA and t is given by $F = t^2$.
- Factorial ANOVA is used when the experimenter wants to study the interaction effects among the treatments.
- Repeated measures ANOVA is used when the same subjects are used for each treatment (e.g., in a longitudinal study).
- Multivariate analysis of variance (MANOVA) is used when there is more than one response variable.

ANOVA cautions

Balanced experiments (those with an equal sample size for each treatment) are relatively easy to interpret; Unbalanced experiments offer more complexity. For single factor (one way) ANOVA, the adjustment for unbalanced data is easy, but the unbalanced analysis lacks both robustness and power.^[52] For more complex designs the lack of balance leads to further complications. "The orthogonality property of main effects and interactions present in balanced data does not carry over to the unbalanced case. This means that the usual analysis of variance techniques do not apply. Consequently, the analysis of unbalanced factorials is much more difficult than that for balanced designs."^[53] In the general case, "The analysis of variance can also be applied to unbalanced data, but then the sums of squares, mean squares, and F-ratios will depend on the order in which the sources of variation are considered." The simplest techniques for handling unbalanced data restore balance by either throwing out data or by synthesizing missing data. More complex techniques use regression.

ANOVA is (in part) a significance test. The American Psychological Association holds the view that simply reporting significance is insufficient and that reporting confidence bounds is preferred.

While ANOVA is conservative (in maintaining a significance level) against multiple comparisons in one dimension, it is not conservative against comparisons in multiple dimensions.^[54]

Generalizations

ANOVA is considered to be a special case of linear regression^{[55][56]} which in turn is a special case of the general linear model.^[57] All consider the observations to be the sum of a model (fit) and a residual (error) to be minimized.

The Kruskal–Wallis test and the Friedman test are nonparametric tests, which do not rely on an assumption of normality.^{[58][59]}

History

While the analysis of variance reached fruition in the 20th century, antecedents extend centuries into the past according to Stigler.^[60] These include hypothesis testing, the partitioning of sums of squares, experimental techniques and the additive model. Laplace was performing hypothesis testing in the 1770s.^[61] The development of least-squares methods by Laplace and Gauss circa 1800 provided an improved method of combining observations (over the existing practices of astronomy and geodesy). It also initiated much study of the contributions to sums of squares. Laplace soon knew how to estimate a variance from a residual (rather than a total) sum of squares.^[62] By 1827 Laplace was using least squares methods to address ANOVA problems regarding measurements of atmospheric tides.^[63] Before 1800 astronomers had isolated observational errors resulting from reaction times (the "personal equation") and had developed methods of reducing the errors.^[64] The experimental methods used in the study of the personal equation were later accepted by the emerging field of psychology^[65] which developed strong (full factorial) experimental methods to which randomization and blinding were soon added.^[66] An eloquent non-mathematical explanation of the additive effects model was available in 1885.^[67]

Sir Ronald Fisher introduced the term "variance" and proposed a formal analysis of variance in a 1918 article *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*.^[68] His first application of the analysis of variance was published in 1921.^[69] Analysis of variance became widely known after being included in Fisher's 1925 book *Statistical Methods for Research Workers*.

Randomization models were developed by several researchers. The first was published in Polish by Neyman in 1923.^[70]

One of the attributes of ANOVA which ensured its early popularity was computational elegance. The structure of the additive model allows solution for the additive coefficients by simple algebra rather than by matrix calculations. In the era of mechanical calculators this simplicity was critical. The determination of statistical significance also required access to tables of the F function which were supplied by early statistics texts.

Footnotes

- [1] Gelman (2005, p 2)
- [2] Howell (2002, p 320)
- [3] Montgomery (2001, p 63)
- [4] Gelman (2005, p 1)
- [5] Gelman (2005, p 5)
- [6] Randomization is a term used in multiple ways in this material. "Randomization has three roles in applications: as a device for eliminating biases, for example from unobserved explanatory variables and selection effects: as a basis for estimating standard errors: and as a foundation for formally exact significance tests." Cox (2006, page 192) Hinkelmann and Kempthorne use randomization both in experimental design and for statistical analysis.
- [7] Montgomery (2001, Chapter 12: Experiments with random factors)
- [8] Gelman (2005, pp 20–21)
- [9] Cochran & Cox (1992, p 48)
- [10] Howell (2002, p 323)
- [11] Anscombe (1948)
- [12] Unit-treatment additivity is simply termed additivity in most texts. Hinkelmann and Kempthorne add adjectives and distinguish between additivity in the strict and broad senses. This allows a detailed consideration of multiple error sources (treatment, state, selection, measurement and sampling) on page 161.
- [13] Kempthorne (1979, p 30)
- [14] Cox (1958, Chapter 2: Some Key Assumptions)
- [15] Hinkelmann and Kempthorne (2008, Volume 1, Throughout. Introduced in Section 2.3.3: Principles of experimental design; The linear model; Outline of a model)
- [16] Hinkelmann and Kempthorne (2008, Volume 1, Section 6.3: Completely Randomized Design; Derived Linear Model)
- [17] Hinkelmann and Kempthorne (2008, Volume 1, Section 6.6: Completely randomized design; Approximating the randomization test)
- [18] Bailey (2008, Chapter 2.14 "A More General Model" in Bailey, pp. 38–40)
- [19] Hinkelmann and Kempthorne (2008, Volume 1, Chapter 7: Comparison of Treatments)
- [20] Kempthorne (1979, pp 125–126, "The experimenter must decide which of the various causes that he feels will produce variations in his results must be controlled experimentally. Those causes that he does not control experimentally, because he is not cognizant of them, he must control by the device of randomization." "[O]nly when the treatments in the experiment are applied by the experimenter using the full randomization procedure is the chain of inductive inference sound. It is *only* under these circumstances that the experimenter can attribute whatever effects he observes to the treatment and the treatment only. Under these circumstances his conclusions are reliable in the statistical sense.")
- [21] Freedman
- [22] Montgomery (2001, Section 3.8: Discovering dispersion effects)
- [23] Hinkelmann and Kempthorne (2008, Volume 1, Section 6.10: Completely randomized design; Transformations)
- [24] Bailey (2008)
- [25] Montgomery (2001, Section 3-3: Experiments with a single factor: The analysis of variance; Analysis of the fixed effects model)
- [26] Cochran & Cox (1992, p 2 example)
- [27] Cochran & Cox (1992, p 49)
- [28] Hinkelmann and Kempthorne (2008, Volume 1, Section 6.7: Completely randomized design; CRD with unequal numbers of replications)
- [29] Rosenbaum (2002, page 40) cites Section 5.7 (Permutation Tests), Theorem 2.3 (actually Theorem 3, page 184) of Lehmann's *Testing Statistical Hypotheses* (1959).
- [30] Moore and McCabe (2003, page 763)
- [31] The F-test for the comparison of variances has a mixed reputation. It is not recommended as a hypothesis test to determine whether two *different* samples have the same variance. It is recommended for ANOVA where two estimates of the variance of the *same* sample are compared. While the F-test is not generally robust against departures from normality, it has been found to be robust in the special case of ANOVA. Citations from Moore & McCabe (2003): "Analysis of variance uses F statistics, but these are not the same as the F statistic for comparing two population standard deviations." (page 554) "The F test and other procedures for inference about variances are so lacking in robustness as to be of little use in practice." (page 556) "[The ANOVA F test] is relatively insensitive to moderate nonnormality and unequal variances, especially when the sample sizes are similar." (page 763) ANOVA assumes homoscedasticity, but it is robust. The statistical test for homoscedasticity (the F-test) is not robust. Moore & McCabe recommend a rule of thumb.
- [32] Gelman (2008)
- [33] Montgomery (2001, Section 5-2: Introduction to factorial designs; The advantages of factorials)
- [34] Belle (2008, Section 8.4: High-order interactions occur rarely)
- [35] Montgomery (2001, Section 5-1: Introduction to factorial designs; Basic definitions and principles)
- [36] Cox (1958, Chapter 6: Basic ideas about factorial experiments)
- [37] Montgomery (2001, Section 5-3.7: Introduction to factorial designs; The two-factor factorial design; One observation per cell)
- [38] Wilkinson (1999, p 596)

- [39] Montgomery (2001, Section 3-7: Determining sample size)
- [40] Howell (2002, Chapter 8: Power)
- [41] Howell (2002, Section 11.12: Power (in ANOVA))
- [42] Howell (2002, Section 13.7: Power analysis for factorial experiments)
- [43] Moore and McCabe (2003, pp 778–780)
- [44] Wilkinson (1999, p 599)
- [45] Montgomery (2001, Section 3-4: Model adequacy checking)
- [46] Moore and McCabe (2003, p 755, Qualifications to this rule appear in a footnote.)
- [47] Montgomery (2001, Section 3-5.8: Experiments with a single factor: The analysis of variance; Practical interpretation of results; Comparing means with a control)
- [48] Hinkelmann and Kempthorne (2008, Volume 1, Section 7.5: Comparison of Treatments; Multiple Comparison Procedures)
- [49] Howell (2002, Chapter 12: Multiple comparisons among treatment means)
- [50] Montgomery (2001, Section 3-5: Practical interpretation of results)
- [51] Cochran & Cox (1957, p 9, "[T]he general rule [is] that the way in which the experiment is conducted determines not only whether inferences can be made, but also the calculations required to make them.")
- [52] Montgomery (2001, Section 3-3.4: Unbalanced data)
- [53] Montgomery (2001, Section 14-2: Unbalanced data in factorial design)
- [54] Wilkinson (1999, p 600)
- [55] Gelman (2005, p.1) (with qualification in the later text)
- [56] Montgomery (2001, Section 3.9: The Regression Approach to the Analysis of Variance)
- [57] Howell (2002, p 604)
- [58] Howell (2002, Chapter 18: Resampling and nonparametric approaches to data)
- [59] Montgomery (2001, Section 3-10: Nonparametric methods in the analysis of variance)
- [60] Stigler (1986)
- [61] Stigler (1986, p 134)
- [62] Stigler (1986, p 153)
- [63] Stigler (1986, pp 154–155)
- [64] Stigler (1986, pp 240–242)
- [65] Stigler (1986, Chapter 7 - Psychophysics as a Counterpoint)
- [66] Stigler (1986, p 253)
- [67] Stigler (1986, pp 314–315)
- [68] *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*. Ronald A. Fisher. *Philosophical Transactions of the Royal Society of Edinburgh*. 1918. (volume 52, pages 399–433)
- [69] On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Ronald A. Fisher. *Metron*, 1: 3-32 (1921)
- [70] Scheffé (1959, p 291, "Randomization models were first formulated by Neyman (1923) for the completely randomized design, by Neyman (1935) for randomized blocks, by Welch (1937) and Pitman (1937) for the Latin square under a certain null hypothesis, and by Kempthorne (1952, 1955) and Wilk (1955) for many other designs.")

Notes

References

- Anscombe, F. J. (1948). "The Validity of Comparative Experiments". *Journal of the Royal Statistical Society. Series A (General)* **111** (3): 181–211. doi: 10.2307/2984159 (<http://dx.doi.org/10.2307/2984159>). JSTOR 2984159 (<http://www.jstor.org/stable/2984159>). MR 30181 (<http://www.ams.org/mathscinet-getitem?mr=30181>).
- Bailey, R. A. (2008). *Design of Comparative Experiments* (<http://www.maths.qmul.ac.uk/~rab/DOEbook>). Cambridge University Press. ISBN 978-0-521-68357-9. Pre-publication chapters are available on-line.
- Belle, Gerald van (2008). *Statistical rules of thumb* (2nd ed.). Hoboken, N.J: Wiley. ISBN 978-0-470-14448-0.
- Cochran, William G.; Cox, Gertrude M. (1992). *Experimental designs* (2nd ed.). New York: Wiley. ISBN 978-0-471-54567-5.
- Cohen, Jacob (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Routledge ISBN 978-0-8058-0283-2
- Cohen, Jacob (1992). "Statistics a power primer". *Psychology Bulletin* **112** (1): 155–159. doi: 10.1037/0033-2909.112.1.155 (<http://dx.doi.org/10.1037/0033-2909.112.1.155>). PMID 19565683 (<http://>

- www.ncbi.nlm.nih.gov/pubmed/19565683).
- Cox, David R. (1958). *Planning of experiments*. Reprinted as ISBN 978-0-471-57429-3
 - Cox, D. R. (2006). *Principles of statistical inference*. Cambridge New York: Cambridge University Press. ISBN 978-0-521-68567-2.
 - Freedman, David A. (2005). *Statistical Models: Theory and Practice*, Cambridge University Press. ISBN 978-0-521-67105-7
 - Gelman, Andrew (2005). "Analysis of variance? Why it is more important than ever". *The Annals of Statistics* **33**: 1–53. doi: 10.1214/009053604000001048 (<http://dx.doi.org/10.1214/009053604000001048>).
 - Gelman, Andrew (2008). "Variance, analysis of". *The new Palgrave dictionary of economics* (2nd ed.). Basingstoke, Hampshire New York: Palgrave Macmillan. ISBN 978-0-333-78676-5.
 - Hinkelmann, Klaus & Kempthorne, Oscar (2008). *Design and Analysis of Experiments*. I and II (Second ed.). Wiley. ISBN 978-0-470-38551-7.
 - Howell, David C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury/Thomson Learning. ISBN 0-534-37770-X.
 - Kempthorne, Oscar (1979). *The Design and Analysis of Experiments* (Corrected reprint of (1952) Wiley ed.). Robert E. Krieger. ISBN 0-88275-105-0.
 - Lehmann, E.L. (1959) *Testing Statistical Hypotheses*. John Wiley & Sons.
 - Montgomery, Douglas C. (2001). *Design and Analysis of Experiments* (5th ed.). New York: Wiley. ISBN 978-0-471-31649-7.
 - Moore, David S. & McCabe, George P. (2003). *Introduction to the Practice of Statistics* (4e). W H Freeman & Co. ISBN 0-7167-9657-0
 - Rosenbaum, Paul R. (2002). *Observational Studies* (2nd ed.). New York: Springer-Verlag. ISBN 978-0-387-98967-9
 - Scheffé, Henry (1959). *The Analysis of Variance*. New York: Wiley.
 - Stigler, Stephen M. (1986). *The history of statistics : the measurement of uncertainty before 1900*. Cambridge, Mass: Belknap Press of Harvard University Press. ISBN 0-674-40340-1.
 - Wilkinson, Leland (1999). "Statistical Methods in Psychology Journals; Guidelines and Explanations". *American Psychologist* **54** (8): 594–604. doi: 10.1037/0003-066X.54.8.594 (<http://dx.doi.org/10.1037/0003-066X.54.8.594>).

Further reading

- Box, G. E. P. (1953). "Non-Normality and Tests on Variances". *Biometrika* (Biometrika Trust) **40** (3/4): 318–335. JSTOR 2333350 (<http://www.jstor.org/stable/2333350>).
- Box, G. E. P. (1954). "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification". *The Annals of Mathematical Statistics* **25** (2): 290. doi: 10.1214/aoms/1177728786 (<http://dx.doi.org/10.1214/aoms/1177728786>).
- Box, G. E. P. (1954). "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, II. Effects of Inequality of Variance and of Correlation Between Errors in the Two-Way Classification". *The Annals of Mathematical Statistics* **25** (3): 484. doi: 10.1214/aoms/1177728717 (<http://dx.doi.org/10.1214/aoms/1177728717>).
- Caliński, Tadeusz & Kageyama, Sanpei (2000). *Block designs: A Randomization approach, Volume I: Analysis*. Lecture Notes in Statistics **150**. New York: Springer-Verlag. ISBN 0-387-98578-6.
- Christensen, Ronald (2002). *Plane Answers to Complex Questions: The Theory of Linear Models* (Third ed.). New York: Springer. ISBN 0-387-95361-2.
- Cox, David R. & Reid, Nancy M. (2000). *The theory of design of experiments*. (Chapman & Hall/CRC). ISBN 978-1-58488-195-7

- Fisher, Ronald (1918). "Studies in Crop Variation. I. An examination of the yield of dressed grain from Broadbalk" (<http://www.library.adelaide.edu.au/digitised/fisher/15.pdf>). *Journal of Agricultural Science* **11**: 107–135. doi: 10.1017/S0021859600003750 (<http://dx.doi.org/10.1017/S0021859600003750>).
- Freedman, David A.; Pisani, Robert; Purves, Roger (2007) *Statistics*, 4th edition. W.W. Norton & Company ISBN 978-0-393-92972-0
- Hettmansperger, T. P.; McKean, J. W. (1998). Edward Arnold, ed. *Robust nonparametric statistical methods*. Kendall's Library of Statistics. Volume 5 (First ed.). New York: John Wiley & Sons, Inc. pp. xiv+467 pp. ISBN 0-340-54937-8. MR 1604954 (<http://www.ams.org/mathscinet-getitem?mr=1604954>).
- Lentner, Marvin; Thomas Bishop (1993). *Experimental design and analysis* (Second ed.). P.O. Box 884, Blacksburg, VA 24063: Valley Book Company. ISBN 0-9616255-2-X.
- Tabachnick, Barbara G. & Fidell, Linda S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Pearson International Edition. ISBN 978-0-205-45938-4
- Wichura, Michael J. (2006). *The coordinate-free approach to linear models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. pp. xiv+199. ISBN 978-0-521-86842-6. MR 2283455 (<http://www.ams.org/mathscinet-getitem?mr=2283455>).

External links

- SOCR ANOVA Activity (http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_ANOVA_1Way) and interactive applet (http://www.socr.ucla.edu/htmls/ana/ANOVA1Way_Analysis.html).
- Examples of all ANOVA and ANCOVA models with up to three treatment factors, including randomized block, split plot, repeated measures, and Latin squares, and their analysis in R (<http://www.southampton.ac.uk/~cpd/anovas/datasets/index.htm>)
- NIST/SEMATECH e-Handbook of Statistical Methods, section 7.4.3: "Are the means equal?" (<http://www.itl.nist.gov/div898/handbook/prc/section4/prc43.htm>)

Principal component analysis

Principal component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components**. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables.

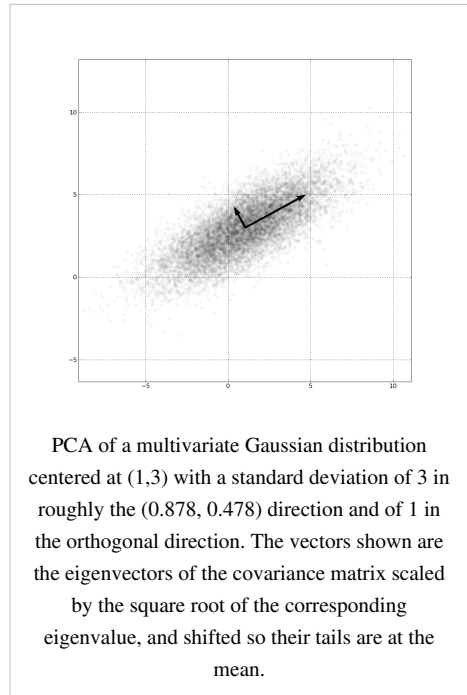
Depending on the field of application, it is also named the discrete Karhunen–Loève transform (KLT) in signal processing, the Hotelling transform in multivariate quality control, proper orthogonal decomposition (POD) in mechanical engineering, singular value decomposition (SVD) of \mathbf{X} (Golub and Van Loan, 1983), eigenvalue decomposition (EVD) of $\mathbf{X}^T\mathbf{X}$ in linear algebra, factor analysis, Eckart–Young theorem (Harman, 1960), or Schmidt–Mirsky theorem in psychometrics, empirical orthogonal functions (EOF) in meteorological science, empirical eigenfunction decomposition (Sirovich, 1987), empirical component analysis (Lorenz, 1956), quasiharmonic modes (Brooks et al., 1988), spectral decomposition in noise and vibration, and empirical modal analysis in structural dynamics.

PCA was invented in 1901 by Karl Pearson, as an analogue of the principal axes theorem in mechanics; it was later independently developed (and named) by Harold Hotelling in the 1930s.^[1] The method is mostly used as a tool in exploratory data analysis and for making predictive models. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z-scores) the data matrix for each attribute. The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point), and loadings (the weight by which each standardized original variable should be multiplied to get the component score).^[2]

PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualised as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a projection or "shadow" of this object when viewed from its (in some sense; see below) most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

PCA is closely related to factor analysis. Factor analysis typically incorporates more domain specific assumptions about the underlying structure and solves eigenvectors of a slightly different matrix.

PCA is also related to canonical correlation analysis (CCA). CCA defines coordinate systems that optimally describe the cross-covariance between two datasets while PCA defines a new orthogonal coordinate system that optimally describes variance in a single dataset.



Details

PCA is mathematically defined^[3] as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

Consider a data matrix, \mathbf{X} , with zero empirical mean (the empirical (sample) mean of the distribution has been subtracted from the data set), where each of the n rows represents a different repetition of the experiment, and each of the p columns gives a particular kind of datum (say, the results from a particular sensor).

Mathematically, the transformation is defined by a set of p -dimensional vectors of weights or *loadings* $\mathbf{w}_{(k)} = (w_1, \dots, w_p)_{(k)}$ that map each row vector $\mathbf{x}_{(i)}$ of \mathbf{X} to a new vector of principal component *scores* $\mathbf{t}_{(i)} = (t_1, \dots, t_p)_{(i)}$, given by

$$t_{k(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)}$$

in such a way that the individual variables of \mathbf{t} considered over the data set successively inherit the maximum possible variance from \mathbf{x} , with each loading vector \mathbf{w} constrained to be a unit vector.

First component

The first loading vector $\mathbf{w}_{(1)}$ thus has to satisfy

$$\mathbf{w}_{(1)} = \arg \max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)_{(i)}^2 \right\} = \arg \max_{\|\mathbf{w}\|=1} \sum_i (\mathbf{x}_{(i)} \cdot \mathbf{w})^2$$

Equivalently, writing this in matrix form gives

$$\mathbf{w}_{(1)} = \arg \max_{\|\mathbf{w}\|=1} \left\{ \|\mathbf{X}\mathbf{w}\|^2 \right\} = \arg \max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right\}$$

Since $\mathbf{w}_{(1)}$ has been defined to be a unit vector, it equivalently also satisfies

$$\mathbf{w}_{(1)} = \arg \max_{\|\mathbf{w}\|=1} \left\{ \frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

The quantity to be maximised can be recognised as a Rayleigh quotient. A standard result for a symmetric matrix such as $\mathbf{X}^T \mathbf{X}$ is that the quotient's maximum possible value is the largest eigenvalue of the matrix, which occurs when \mathbf{w} is the corresponding eigenvector.

With $\mathbf{w}_{(1)}$ found, the first component of a data vector $\mathbf{x}_{(i)}$ can then be given as a score $t_{1(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}$ in the transformed co-ordinates, or as the corresponding vector in the original variables, $\{\mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}\} \mathbf{w}_{(1)}$.

Further components

The k th component can be found by subtracting the first $k - 1$ principal components from \mathbf{X} :

$$\hat{\mathbf{X}}_{k-1} = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X} \mathbf{w}_{(s)} \mathbf{w}_{(s)}^T$$

and then finding the loading vector which extracts the maximum variance from this new data matrix

$$\mathbf{w}_{(k)} = \arg \max_{\|\mathbf{w}\|=1} \left\{ \|\hat{\mathbf{X}}_{k-1} \mathbf{w}\|^2 \right\} = \arg \max_{\|\mathbf{w}\|=1} \left\{ \frac{\mathbf{w}^T \hat{\mathbf{X}}_{k-1}^T \hat{\mathbf{X}}_{k-1} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

It turns out that this gives the remaining eigenvectors of $\mathbf{X}^T \mathbf{X}$, with the maximum values for the quantity in brackets given by their corresponding eigenvalues.

The k th principal component of a data vector $\mathbf{x}_{(i)}$ can therefore be given as a score $t_{k(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)}$ in the transformed co-ordinates, or as the corresponding vector in the space of the original variables, $\{\mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)}\} \mathbf{w}_{(k)}$, where $\mathbf{w}_{(k)}$ is the k th eigenvector of $\mathbf{X}^T \mathbf{X}$.

The full principal components decomposition of \mathbf{X} can therefore be given as

$$\mathbf{T} = \mathbf{X}\mathbf{W}$$

where \mathbf{W} is a p -by- p matrix whose columns are the eigenvectors of $\mathbf{X}^T\mathbf{X}$

Covariances

$\mathbf{X}^T\mathbf{X}$ itself can be recognised as proportional to the empirical sample covariance matrix of the dataset \mathbf{X} .

The sample covariance Q between two of the different principal components over the dataset is given by

$$\begin{aligned} Q(\text{PC}_{(j)}, \text{PC}_{(k)}) &\propto (\mathbf{X}\mathbf{w}_{(j)}) \cdot (\mathbf{X}\mathbf{w}_{(k)}) \\ &= \mathbf{w}_{(j)}^T \mathbf{X}^T \mathbf{X} \mathbf{w}_{(k)} \\ &= \mathbf{w}_{(j)}^T \lambda_{(k)} \mathbf{w}_{(k)} \\ &= \lambda_{(k)} \mathbf{w}_{(j)}^T \mathbf{w}_{(k)} \end{aligned}$$

where the eigenvector property of $\mathbf{w}_{(k)}$ has been used to move from line 2 to line 3. However eigenvectors $\mathbf{w}_{(j)}$ and $\mathbf{w}_{(k)}$ corresponding to eigenvalues of a symmetric matrix are orthogonal (if the eigenvalues are different), or can be orthogonalised (if the vectors happen to share an equal repeated value). The product in the final line is therefore zero; there is no sample covariance between different principal components over the dataset.

Another way to characterise the principal components transformation is therefore as the transformation to coordinates which diagonalise the empirical sample covariance matrix.

In matrix form, the empirical covariance matrix for the original variables can be written

$$\mathbf{Q} \propto \mathbf{X}^T \mathbf{X} = \mathbf{W} \mathbf{\Lambda} \mathbf{W}^T$$

The empirical covariance matrix between the principal components becomes

$$\mathbf{W}^T \mathbf{Q} \mathbf{W} \propto \mathbf{W}^T \mathbf{W} \mathbf{\Lambda} \mathbf{W}^T \mathbf{W} = \mathbf{\Lambda}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues $\lambda_{(k)}$ of $\mathbf{X}^T\mathbf{X}$

($\lambda_{(k)}$ being equal to the sum of the squares over the dataset associated with each component k : $\lambda_{(k)} = \sum_i t_{k(i)}^2 = \sum_i (\mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)})^2$)

Dimensionality reduction

The faithful transformation $\mathbf{T} = \mathbf{X}\mathbf{W}$ maps a data vector $\mathbf{x}_{(i)}$ from an original space of p variables to a new space of p variables which are uncorrelated over the dataset. However, not all the principal components need to be kept. Keeping only the first L principal components, produced by using only the first L loading vectors, gives the truncated transformation

$$\mathbf{T}_L = \mathbf{X}\mathbf{W}_L$$

where the matrix \mathbf{T}_L now has n rows but only L columns. By construction, of all the transformed data matrices with only L columns, this score matrix maximises the variance in the original data that has been preserved, while minimising the total squared reconstruction error $\|\mathbf{T} - \mathbf{T}_L\|^2$.

Such dimensionality reduction can be a very useful step for visualising and processing high-dimensional datasets, while still retaining as much of the variance in the dataset as possible. For example, selecting $L = 2$ and keeping only the first two principal components finds the two-dimensional plane through the high-dimensional dataset in which the data is most spread out, so if the data contains clusters these too may be most spread out, and therefore most visible to be plotted out in a two-dimensional diagram; whereas if two directions through the data (or two of the original variables) are chosen at random, the clusters may be much less spread apart from each other, and may in fact be much more likely to substantially overlay each other, making them indistinguishable.

Similarly, in regression analysis, the larger the number of explanatory variables allowed, the greater is the chance of overfitting the model, producing conclusions that fail to generalise to other datasets. One approach, especially when there are strong correlations between different possible explanatory variables, is to reduce them to a few principal components and then run the regression against them, a method called principal component regression.

Dimensionality reduction may also be appropriate when the variables in a dataset are noisy. If each column of the dataset contains independent identically distributed Gaussian noise, then the columns of \mathbf{T} will also contain similarly identically distributed Gaussian noise (such a distribution is invariant under the effects of the matrix \mathbf{W} , which can be thought of as a high-dimensional rotation of the co-ordinate axes). However, with more of the total variance concentrated in the first few principal components compared to the same noise variance, the proportionate effect of the noise is less—the first components achieve a higher signal-to-noise ratio. PCA thus can have the effect of concentrating much of the signal into the first few principal components, which can usefully be captured by dimensionality reduction; while the later principal components may be dominated by noise, and so disposed of without great loss.

Singular value decomposition

The principal components transformation can also be associated with another matrix factorisation, the singular value decomposition (SVD) of \mathbf{X} ,

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T$$

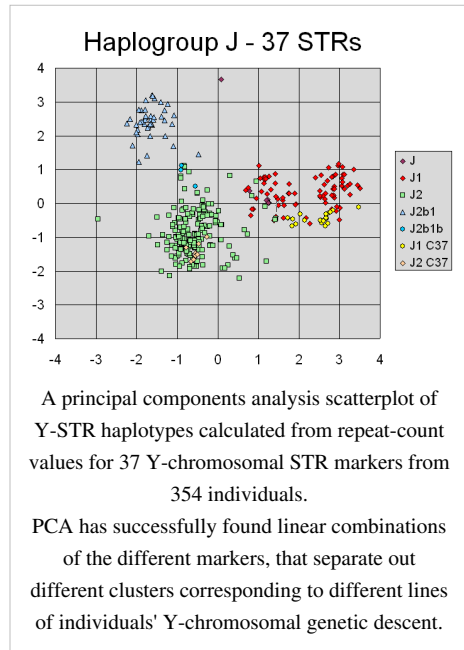
Here $\mathbf{\Sigma}$ is a n -by- p rectangular diagonal matrix of positive numbers $\sigma_{(k)}$, called the singular values of \mathbf{X} ; \mathbf{U} is an n -by- n matrix, the columns of which are orthogonal unit vectors of length n called the left singular vectors of \mathbf{X} ; and \mathbf{W} is a p -by- p whose columns are orthogonal unit vectors of length p and called the right singular vectors of \mathbf{X} .

In terms of this factorisation, the matrix $\mathbf{X}^T\mathbf{X}$ can be written

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= \mathbf{W}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{W}^T \\ &= \mathbf{W}\mathbf{\Sigma}^2\mathbf{W}^T\end{aligned}$$

Comparison with the eigenvector factorisation of $\mathbf{X}^T\mathbf{X}$ establishes that the right singular vectors \mathbf{W} of \mathbf{X} are equivalent to the eigenvectors of $\mathbf{X}^T\mathbf{X}$, while the singular values $\sigma_{(k)}$ of \mathbf{X} are equal to the square roots of the eigenvalues $\lambda_{(k)}$ of $\mathbf{X}^T\mathbf{X}$.

Using the singular value decomposition the score matrix \mathbf{T} can be written



$$\begin{aligned}\mathbf{T} &= \mathbf{X}\mathbf{W} \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T\mathbf{W} \\ &= \mathbf{U}\mathbf{\Sigma}\end{aligned}$$

so each column of \mathbf{T} is given by one of the left singular vectors of \mathbf{X} multiplied by the corresponding singular value.

Efficient algorithms exist to calculate the SVD of \mathbf{X} without having to form the matrix $\mathbf{X}^T\mathbf{X}$, so computing the SVD is now the standard way to calculate a principal components analysis from a data matrix, unless only a handful of components are required.

As with the eigendecomposition, a truncated n -by- L score matrix \mathbf{T}_L can be obtained by considering only the first L largest singular values and their singular vectors:

$$\mathbf{T}_L = \mathbf{U}_L\mathbf{\Sigma}_L = \mathbf{X}\mathbf{W}_L$$

The truncation of a matrix \mathbf{M} or \mathbf{T} using a truncated singular value decomposition in this way produces a truncated matrix that is the nearest possible matrix of rank L to the original matrix, in the sense of the difference between the two having the smallest possible Frobenius norm, a result known as the Eckart–Young theorem [1936].

Further considerations

Given a set of points in Euclidean space, the first principal component corresponds to a line that passes through the multidimensional mean and minimizes the sum of squares of the distances of the points from the line. The second principal component corresponds to the same concept after all correlation with the first principal component has been subtracted from the points. The singular values (in $\mathbf{\Sigma}$) are the square roots of the eigenvalues of the matrix $\mathbf{X}^T\mathbf{X}$. Each eigenvalue is proportional to the portion of the "variance" (more correctly of the sum of the squared distances of the points from their multidimensional mean) that is correlated with each eigenvector. The sum of all the eigenvalues is equal to the sum of the squared distances of the points from their multidimensional mean. PCA essentially rotates the set of points around their mean in order to align with the principal components. This moves as much of the variance as possible (using an orthogonal transformation) into the first few dimensions. The values in the remaining dimensions, therefore, tend to be small and may be dropped with minimal loss of information (see below). PCA is often used in this manner for dimensionality reduction. PCA has the distinction of being the optimal orthogonal transformation for keeping the subspace that has largest "variance" (as defined above). This advantage, however, comes at the price of greater computational requirements if compared, for example and when applicable, to the discrete cosine transform, and in particular to the DCT-II which is simply known as the "DCT". Nonlinear dimensionality reduction techniques tend to be more computationally demanding than PCA.

PCA is sensitive to the scaling of the variables. If we have just two variables and they have the same sample variance and are positively correlated, then the PCA will entail a rotation by 45° and the "loadings" for the two variables with respect to the principal component will be equal. But if we multiply all values of the first variable by 100, then the principal component will be almost the same as that variable, with a small contribution from the other variable, whereas the second component will be almost aligned with the second original variable. This means that whenever the different variables have different units (like temperature and mass), PCA is a somewhat arbitrary method of analysis. (Different results would be obtained if one used Fahrenheit rather than Celsius for example.) Note that Pearson's original paper was entitled "On Lines and Planes of Closest Fit to Systems of Points in Space" – "in space" implies physical Euclidean space where such concerns do not arise. One way of making the PCA less arbitrary is to use variables scaled so as to have unit variance, by standardizing the data and hence use the autocorrelation matrix instead of the autocovariance matrix as a basis for PCA. However, this compresses the fluctuations in all dimensions of the signal space to unit variance.

Mean subtraction (a.k.a. "mean centering") is necessary for performing PCA to ensure that the first principal component describes the direction of maximum variance. If mean subtraction is not performed, the first principal component might instead correspond more or less to the mean of the data. A mean of zero is needed for finding a

basis that minimizes the mean square error of the approximation of the data.^[4]

PCA is equivalent to empirical orthogonal functions (EOF), a name which is used in meteorology.

An autoencoder neural network with a linear hidden layer is similar to PCA. Upon convergence, the weight vectors of the K neurons in the hidden layer will form a basis for the space spanned by the first K principal components. Unlike PCA, this technique will not necessarily produce orthogonal vectors.

PCA is a popular primary technique in pattern recognition. It is not, however, optimized for class separability. An alternative is the linear discriminant analysis, which does take this into account.

Another application of PCA is reducing the number of parameters in the process of generating computational models of oil reservoirs.^[5]

Table of symbols and abbreviations

Symbol	Meaning	Dimensions	Indices
$\mathbf{X} = \{X[i, j]\}$	data matrix, consisting of the set of all data vectors, one vector per row	$n \times p$	$i = 1 \dots n$ $j = 1 \dots p$
n	the number of row vectors in the data set	1×1	scalar
p	the number of elements in each row vector (dimension)	1×1	scalar
L	the number of dimensions in the dimensionally reduced subspace, $1 \leq L \leq p$	1×1	scalar
$\mathbf{u} = \{u[j]\}$	vector of empirical means, one mean for each column j of the data matrix	$p \times 1$	$j = 1 \dots p$
$\mathbf{s} = \{s[j]\}$	vector of empirical standard deviations, one standard deviation for each column j of the data matrix	$p \times 1$	$j = 1 \dots p$
$\mathbf{h} = \{h[i]\}$	vector of all 1's	$1 \times n$	$i = 1 \dots n$
$\mathbf{B} = \{B[i, j]\}$	deviations from the mean of each column j of the data matrix	$n \times p$	$i = 1 \dots n$ $j = 1 \dots p$
$\mathbf{Z} = \{Z[m, n]\}$	z-scores, computed using the mean and standard deviation for each row m of the data matrix	$n \times p$	$i = 1 \dots n$ $j = 1 \dots p$
$\mathbf{C} = \{C[k, l]\}$	covariance matrix	$p \times p$	$k = 1 \dots p$ $l = 1 \dots p$
$\mathbf{R} = \{R[k, l]\}$	correlation matrix	$p \times p$	$k = 1 \dots p$ $l = 1 \dots p$
$\mathbf{V} = \{V[j, k]\}$	matrix consisting of the set of all eigenvectors of \mathbf{C} , one eigenvector per column	$p \times p$	$j = 1 \dots p$ $k = 1 \dots p$
$\mathbf{D} = \{D[k, l]\}$	diagonal matrix consisting of the set of all eigenvalues of \mathbf{C} along its principal diagonal, and 0 for all other elements	$p \times p$	$k = 1 \dots p$ $l = 1 \dots p$
$\mathbf{W} = \{W[j, k]\}$	matrix of basis vectors, one vector per column, where each basis vector is one of the eigenvectors of \mathbf{C} , and where the vectors in \mathbf{W} are a sub-set of those in \mathbf{V}	$p \times L$	$j = 1 \dots p$ $k = 1 \dots L$
$\mathbf{T} = \{T[i, k]\}$	matrix consisting of n row vectors, where each vector is the projection of the corresponding data vector from matrix \mathbf{X} onto the basis vectors contained in the columns of matrix \mathbf{W} .	$n \times L$	$i = 1 \dots n$ $k = 1 \dots L$

Properties and limitations of PCA

Properties ^[6]

Property 1: For any integer q , $1 \leq q \leq p$, consider the orthogonal linear transformation

$$\mathbf{y} = \mathbf{B}'\mathbf{x}$$

where \mathbf{y} is a q -element vector and \mathbf{B}' is a $(q \times p)$ matrix, and let $\Sigma_{\mathbf{y}} = \mathbf{B}'\Sigma\mathbf{B}$ be the variance-covariance matrix for \mathbf{y} . Then the trace of $\Sigma_{\mathbf{y}}$, denoted $\text{tr}(\Sigma_{\mathbf{y}})$, is maximized by taking $\mathbf{B} = \mathbf{A}_q$, where \mathbf{A}_q consists of the first q columns of \mathbf{A} (\mathbf{B}' is the transposition of \mathbf{B}).

Property 2: Consider again the orthonormal transformation

$$\mathbf{y} = \mathbf{B}'\mathbf{x}$$

with \mathbf{x} , \mathbf{B} , \mathbf{A} and $\Sigma_{\mathbf{y}}$ defined as before. Then $\text{tr}(\Sigma_{\mathbf{y}})$ is minimized by taking $\mathbf{B} = \mathbf{A}_q^*$, where \mathbf{A}_q^* consists of the last q columns of \mathbf{A} .

The statistical implication of this property is that the last few PCs are not simply unstructured left-overs after removing the important PCs. Because these last PCs have variances as small as possible they are useful in their own right. They can help to detect unsuspected near-constant linear relationships between the elements of \mathbf{x} , and they may also be useful in regression, in selecting a subset of variables from \mathbf{x} , and in outlier detection.

Property 3: (*the Spectral Decomposition of Σ*)

$$\Sigma = \lambda_1\alpha_1\alpha_1' + \lambda_2\alpha_2\alpha_2' + \dots + \lambda_p\alpha_p\alpha_p'$$

Before we look at its usage, we first look at diagonal elements,

$$\text{var}(x_j) = \sum_{k=1}^p \lambda_k \alpha_{kj}^2$$

Then, perhaps the main statistical implication of the result is that not only can we decompose the combined variances of all the elements of \mathbf{x} into decreasing contributions due to each PC, but we can also decompose the whole covariance matrix into contributions $\lambda_k\alpha_k\alpha_k'$ from each PC. Although not strictly decreasing, the elements of $\lambda_k\alpha_k\alpha_k'$ will tend to become smaller as k increases, as $\lambda_k\alpha_k\alpha_k'$ decreases for increasing k , whereas the elements of α_k tend to stay 'about the same size' because of the normalization constraints: $\alpha_k'\alpha_k = 1$, $k = 1, 2, \dots, p$

Limitations

As noted above, the results of PCA depend on the scaling of the variables.

The applicability of PCA is limited by certain assumptions^[7] made in its derivation.

PCA and information theory

The claim that the PCA used for dimensionality reduction preserves most of the information of the data is misleading. Indeed, without any assumption on the signal model, PCA cannot help to reduce the amount of information lost during dimensionality reduction, where information was measured using Shannon entropy.

Under the assumption that

$$\mathbf{x} = \mathbf{s} + \mathbf{n}$$

i.e., that the data vector \mathbf{x} is the sum of the desired information-bearing signal \mathbf{s} and a noise signal \mathbf{n} one can show that PCA can be optimal for dimensionality reduction also from an information-theoretic point-of-view.

In particular, Linsker showed that if \mathbf{s} is Gaussian and \mathbf{n} is Gaussian noise with a covariance matrix proportional to the identity matrix, the PCA maximizes the mutual information $I(\mathbf{y}; \mathbf{s})$ between the desired information \mathbf{s} and the dimensionality-reduced output $\mathbf{y} = \mathbf{W}_L^T \mathbf{x}$.

If the noise is still Gaussian and has a covariance matrix proportional to the identity matrix (i.e., the components of the vector \mathbf{n} are iid), but the information-bearing signal \mathbf{s} is non-Gaussian (which is a common scenario), PCA at least minimizes an upper bound on the *information loss*, which is defined as^[8]

$$I(\mathbf{x}; \mathbf{s}) - I(\mathbf{y}; \mathbf{s}).$$

The optimality of PCA is also preserved if the noise \mathbf{n} is iid and at least more Gaussian (in terms of the Kullback–Leibler divergence) than the information-bearing signal \mathbf{s} . In general, even if the above signal model holds, PCA loses its information-theoretic optimality as soon as the noise \mathbf{n} becomes dependent.

Computing PCA using the covariance method

The following is a detailed description of PCA using the covariance method (see also here^[9]). But note that it is better to use the singular value decomposition (using standard software)^{Wikipedia:Quotations}.

The goal is to transform a given data set \mathbf{X} of dimension p to an alternative data set \mathbf{Y} of smaller dimension L . Equivalently, we are seeking to find the matrix \mathbf{Y} , where \mathbf{Y} is the Karhunen–Loève transform (KLT) of matrix \mathbf{X} :

$$\mathbf{Y} = \text{KLT}\{\mathbf{X}\}$$

Organize the data set

Suppose you have data comprising a set of observations of p variables, and you want to reduce the data so that each observation can be described with only L variables, $L < p$. Suppose further, that the data are arranged as a set of n data vectors $\mathbf{x}_1 \dots \mathbf{x}_n$ with each \mathbf{x}_i representing a single grouped observation of the p variables.

- Write $\mathbf{x}_1 \dots \mathbf{x}_n$ as row vectors, each of which has p columns.
- Place the row vectors into a single matrix \mathbf{X} of dimensions $n \times p$.

Calculate the empirical mean

- Find the empirical mean along each dimension $j = 1, \dots, p$.
- Place the calculated mean values into an empirical mean vector \mathbf{u} of dimensions $p \times 1$.

$$u[j] = \frac{1}{N} \sum_{i=1}^n X[i, j]$$

Calculate the deviations from the mean

Mean subtraction is an integral part of the solution towards finding a principal component basis that minimizes the mean square error of approximating the data.^[10] Hence we proceed by centering the data as follows:

- Subtract the empirical mean vector \mathbf{u} from each row of the data matrix \mathbf{X} .
- Store mean-subtracted data in the $n \times p$ matrix \mathbf{B} .

$$\mathbf{B} = \mathbf{X} - \mathbf{h}\mathbf{u}^T$$

where \mathbf{h} is an $n \times 1$ column vector of all 1s:

$$h[i] = 1 \quad \text{for } i = 1, \dots, n$$

Find the covariance matrix

- Find the $p \times p$ empirical covariance matrix \mathbf{C} from the outer product of matrix \mathbf{B} with itself:

$$\mathbf{C} = \frac{1}{n-1} \mathbf{B}^* \cdot \mathbf{B}$$

where

* is the conjugate transpose operator. Note that if \mathbf{B} consists entirely of real numbers, which is the case in many applications, the "conjugate transpose" is the same as the regular transpose.

- Please note that outer products apply to vectors. For tensor cases we should apply tensor products, but the covariance matrix in PCA is a sum of outer products between its sample vectors; indeed, it could be represented as $\mathbf{B}^* \cdot \mathbf{B}$. See the covariance matrix sections on the discussion page for more information.
- The reasoning behind using $N-1$ instead of N to calculate the covariance is Bessel's correction

Find the eigenvectors and eigenvalues of the covariance matrix

- Compute the matrix \mathbf{V} of eigenvectors which diagonalizes the covariance matrix \mathbf{C} :

$$\mathbf{V}^{-1} \mathbf{C} \mathbf{V} = \mathbf{D}$$

where \mathbf{D} is the diagonal matrix of eigenvalues of \mathbf{C} . This step will typically involve the use of a computer-based algorithm for computing eigenvectors and eigenvalues. These algorithms are readily available as sub-components of most matrix algebra systems, such as R, MATLAB,^{[11][12]} Mathematica,^[13] SciPy, IDL (Interactive Data Language), or GNU Octave as well as OpenCV.

- Matrix \mathbf{D} will take the form of an $M \times M$ diagonal matrix, where

$$D[k, l] = \lambda_k \quad \text{for } k = l = j$$

is the j th eigenvalue of the covariance matrix \mathbf{C} , and

$$D[k, l] = 0 \quad \text{for } k \neq l.$$

- Matrix \mathbf{V} , also of dimension $p \times p$, contains p column vectors, each of length p , which represent the p eigenvectors of the covariance matrix \mathbf{C} .
- The eigenvalues and eigenvectors are ordered and paired. The j th eigenvalue corresponds to the j th eigenvector.

Rearrange the eigenvectors and eigenvalues

- Sort the columns of the eigenvector matrix \mathbf{V} and eigenvalue matrix \mathbf{D} in order of *decreasing* eigenvalue.
- Make sure to maintain the correct pairings between the columns in each matrix.

Compute the cumulative energy content for each eigenvector

- The eigenvalues represent the distribution of the source data's energy. Please clarify among each of the eigenvectors, where the eigenvectors form a basis for the data. The cumulative energy content g for the j th eigenvector is the sum of the energy content across all of the eigenvalues from 1 through j :

$$g[j] = \sum_{k=1}^j D[k, k] \quad \text{for } j = 1, \dots, p \text{ [citation needed]}$$

Select a subset of the eigenvectors as basis vectors

- Save the first L columns of \mathbf{V} as the $p \times L$ matrix \mathbf{W} :

$$W[k, l] = V[k, l] \quad \text{for} \quad k = 1, \dots, p \quad l = 1, \dots, L$$

where

$$1 \leq L \leq p.$$

- Use the vector \mathbf{g} as a guide in choosing an appropriate value for L . The goal is to choose a value of L as small as possible while achieving a reasonably high value of g on a percentage basis. For example, you may want to choose L so that the cumulative energy g is above a certain threshold, like 90 percent. In this case, choose the smallest value of L such that

$$\frac{g[L]}{g[p]} \geq 0.9$$

Convert the source data to z-scores (optional)

- Create an $p \times 1$ empirical standard deviation vector \mathbf{s} from the square root of each element along the main diagonal of the diagonalized covariance matrix \mathbf{C} . (Note, that scaling operations do not commute with the KLT thus we must scale by the variances of the already-decorrelated vector, which is the diagonal of \mathbf{C}):

$$\mathbf{s} = \{s[j]\} = \{\sqrt{C[j, j]}\} \quad \text{for } j = 1, \dots, p$$

- Calculate the $n \times p$ z-score matrix:

$$\mathbf{Z} = \frac{\mathbf{B}}{\mathbf{h} \cdot \mathbf{s}^T} \text{ (divide element-by-element)}$$

- Note: While this step is useful for various applications as it normalizes the data set with respect to its variance, it is not integral part of PCA/KLT

Project the z-scores of the data onto the new basis

- The projected vectors are the columns of the matrix

$$\mathbf{T} = \mathbf{Z} \cdot \mathbf{W} = \text{KLT}\{\mathbf{X}\}.$$

- The rows of matrix \mathbf{T} represent the Karhunen–Loeve transforms (KLT) of the data vectors in the rows of matrix \mathbf{X} .

Derivation of PCA using the covariance method

Let \mathbf{X} be a d -dimensional random vector expressed as column vector. Without loss of generality, assume \mathbf{X} has zero mean.

We want to find $(*)$ a $d \times d$ orthonormal transformation matrix \mathbf{P} so that \mathbf{PX} has a diagonal covariant matrix (*i.e.* \mathbf{PX} is a random vector with all its distinct components pairwise uncorrelated).

A quick computation assuming \mathbf{P} were unitary yields:

$$\begin{aligned} \text{var}(\mathbf{PX}) &= \mathbb{E}[\mathbf{PX} (\mathbf{PX})^\dagger] \\ &= \mathbb{E}[\mathbf{PX} \mathbf{X}^\dagger \mathbf{P}^\dagger] \\ &= \mathbf{P} \mathbb{E}[\mathbf{X} \mathbf{X}^\dagger] \mathbf{P}^\dagger \\ &= \mathbf{P} \text{var}(\mathbf{X}) \mathbf{P}^{-1} \end{aligned}$$

Hence $(*)$ holds if and only if $\text{var}(\mathbf{X})$ were diagonalisable by \mathbf{P} .

This is very constructive, as $\text{var}(\mathbf{X})$ is guaranteed to be a non-negative definite matrix and thus is guaranteed to be diagonalisable by some unitary matrix.

Iterative computation

In practical implementations especially with high dimensional data (large p), the covariance method is rarely used because it is not efficient. One way to compute the first principal component efficiently^[14] is shown in the following pseudo-code, for a data matrix \mathbf{X} with zero mean, without ever computing its covariance matrix

```

r = a random vector of length  $p$ 
do  $c$  times:
    s =  $\mathbf{0}$  (a vector of length  $p$ )
    for each row  $\mathbf{x} \in \mathbf{X}$ 
        s = s +  $(\mathbf{x} \cdot \mathbf{r})\mathbf{x}$ 
    r =  $\frac{\mathbf{s}}{|\mathbf{s}|}$ 
return r

```

This algorithm is simply an efficient way of calculating $\mathbf{X}^T \mathbf{X} \mathbf{r}$, normalizing, and placing the result back in \mathbf{r} (power iteration). It avoids the np^2 operations of calculating the covariance matrix. \mathbf{r} will typically get close to the first principal component of \mathbf{X} within a small number of iterations, c . (The magnitude of \mathbf{s} will be larger after each iteration. Convergence can be detected when it increases by an amount too small for the precision of the machine.)

Subsequent principal components can be computed by subtracting component \mathbf{r} from \mathbf{X} (see Gram–Schmidt) and then repeating this algorithm to find the next principal component. However this simple approach is not numerically stable if more than a small number of principal components are required, because imprecisions in the calculations will additively affect the estimates of subsequent principal components. More advanced methods build on this basic idea, as with the closely related Lanczos algorithm.

One way to compute the eigenvalue that corresponds with each principal component is to measure the difference in mean-squared-distance between the rows and the centroid, before and after subtracting out the principal component. The eigenvalue that corresponds with the component that was removed is equal to this difference.

The NIPALS method

For very high-dimensional datasets, such as those generated in the *omics sciences (e.g., genomics, metabolomics) it is usually only necessary to compute the first few PCs. The non-linear iterative partial least squares (NIPALS) algorithm calculates \mathbf{t}_1 and \mathbf{w}_1^T from \mathbf{X} . The outer product, $\mathbf{t}_1 \mathbf{w}_1^T$ can then be subtracted from \mathbf{X} leaving the residual matrix \mathbf{E}_1 . This can be then used to calculate subsequent PCs. This results in a dramatic reduction in computational time since calculation of the covariance matrix is avoided.

However, for large data matrices, or matrices that have a high degree of column collinearity, NIPALS suffers from loss of orthogonality due to machine precision limitations accumulated in each iteration step.^[15] A Gram–Schmidt (GS) re-orthogonalization algorithm is applied to both the scores and the loadings at each iteration step to eliminate this loss of orthogonality.^[16]

Online/sequential estimation

In an "online" or "streaming" situation with data arriving piece by piece rather than being stored in a single batch, it is useful to make an estimate of the PCA projection that can be updated sequentially. This can be done efficiently, but requires different algorithms.

Relation between PCA and *K*-means clustering

It has been shown recently (2001,2004)^{[17][18]} that the relaxed solution of *K*-means clustering, specified by the cluster indicators, is given by the PCA principal components, and the PCA subspace spanned by the principal directions is identical to the cluster centroid subspace specified by the between-class scatter matrix. Thus PCA automatically projects to the subspace where the global solution of *K*-means clustering lies, and thus facilitates *K*-means clustering to find near-optimal solutions.

Relation between PCA and Factor Analysis ^[19]

Principle components creates variables that are linear combinations of the original variables. The new variables have the property that the variables are all orthogonal. The principle components can be used to find clusters in a set of data. PCA is a variance-focused approach seeking to reproduce the total variable variance, in which components reflect both common and unique variance of the variable. PCA is generally preferred for purposes of data reduction (i.e., translating variable space into optimal factor space) but not when detect the latent construct or factors.

Factor analysis is similar to principle component analysis, in that factor analysis also involves linear combinations of variables. Different from PCA, factor analysis is a correlation-focused approach seeking to reproduce the inter-correlations among variables, in which the factors "represent the common variance of variables, excluding unique variance^[20]". Factor analysis is generally used when the research purpose is detecting data structure (i.e., latent constructs or factors) or causal modeling.

Correspondence analysis

Correspondence analysis (CA) was developed by Jean-Paul Benzécri and is conceptually similar to PCA, but scales the data (which should be non-negative) so that rows and columns are treated equivalently. It is traditionally applied to contingency tables. CA decomposes the chi-squared statistic associated to this table into orthogonal factors. Because CA is a descriptive technique, it can be applied to tables for which the chi-squared statistic is appropriate or not. Several variants of CA are available including detrended correspondence analysis and canonical correspondence analysis. One special extension is multiple correspondence analysis, which may be seen as the counterpart of principal component analysis for categorical data.

Generalizations

Nonlinear generalizations

Most of the modern methods for nonlinear dimensionality reduction find their theoretical and algorithmic roots in PCA or K-means. Pearson's original idea was to take a straight line (or plane) which will be "the best fit" to a set of data points. **Principal curves and manifolds**^[24] give the natural geometric framework for PCA generalization and extend the geometric interpretation of PCA by explicitly constructing an embedded manifold for data approximation, and by encoding using standard geometric projection onto the manifold, as it is illustrated by Fig. See also the elastic map algorithm and principal geodesic analysis. Another popular generalization is kernel PCA, which corresponds to PCA performed in a reproducing kernel Hilbert space associated with a positive definite kernel.

Multilinear generalizations

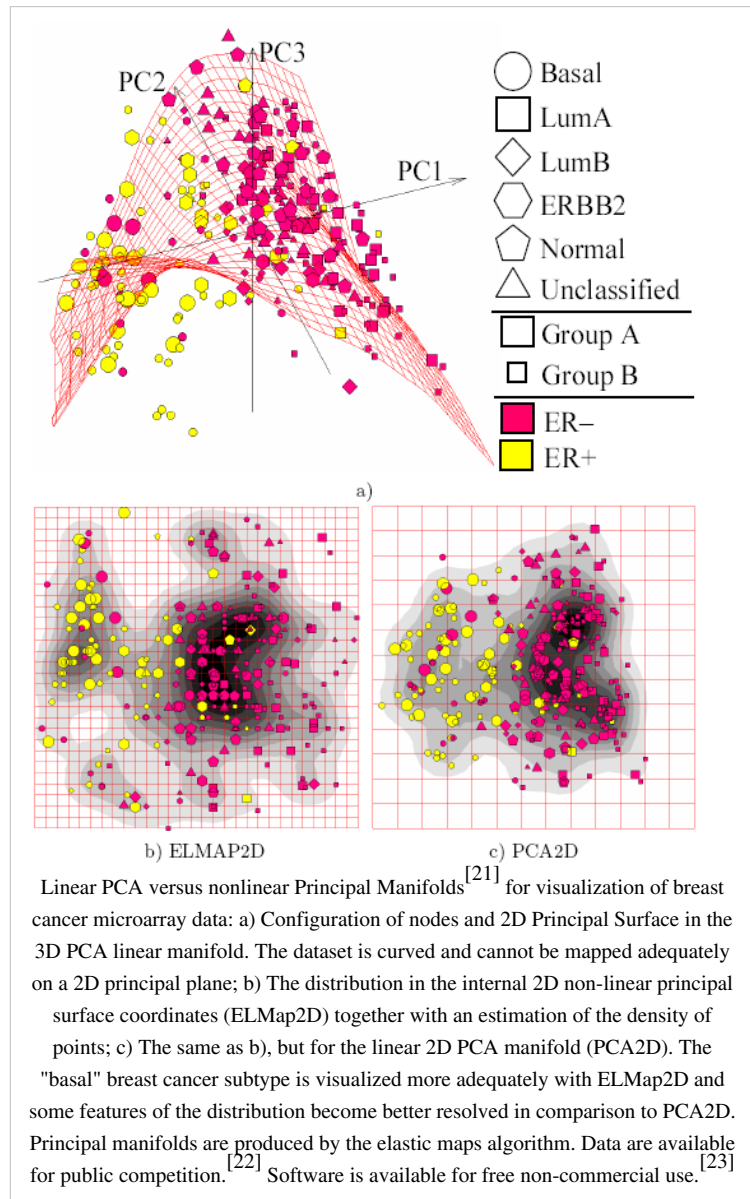
In multilinear subspace learning, PCA is generalized to multilinear PCA (MPCA) that extracts features directly from tensor representations. MPCA is solved by performing PCA in each mode of the tensor iteratively. MPCA has been applied to face recognition, gait recognition, etc. MPCA is further extended to uncorrelated MPCA, non-negative MPCA and robust MPCA.

Higher order

N -way principal component analysis may be performed with models such as Tucker decomposition, PARAFAC, multiple factor analysis, co-inertia analysis, STATIS, and DISTATIS.

Robustness – weighted PCA

While PCA finds the mathematically optimal method (as in minimizing the squared error), it is sensitive to outliers in the data that produce large errors PCA tries to avoid. It therefore is common practice to remove outliers before computing PCA. However, in some contexts, outliers can be difficult to identify. For example in data mining algorithms like correlation clustering, the assignment of points to clusters and outliers is not known beforehand. A



Linear PCA versus nonlinear Principal Manifolds^[21] for visualization of breast cancer microarray data: a) Configuration of nodes and 2D Principal Surface in the 3D PCA linear manifold. The dataset is curved and cannot be mapped adequately on a 2D principal plane; b) The distribution in the internal 2D non-linear principal surface coordinates (ELMap2D) together with an estimation of the density of points; c) The same as b), but for the linear 2D PCA manifold (PCA2D). The "basal" breast cancer subtype is visualized more adequately with ELMap2D and some features of the distribution become better resolved in comparison to PCA2D. Principal manifolds are produced by the elastic maps algorithm. Data are available for public competition.^[22] Software is available for free non-commercial use.^[23]

recently proposed generalization of PCA based on a **weighted PCA** increases robustness by assigning different weights to data objects based on their estimated relevancy.

Software/source code

- Mathematica implements principal component analysis with the `PrincipalComponents` command^[25] using both covariance and correlation methods.
- In the NAG Library, principal components analysis is implemented via the `g03aa` routine (available in both the Fortran and the C versions of the Library).
- In the MATLAB Statistics Toolbox, the functions `princomp` and `pca` (R2012b) give the principal components, while the function `pcares` gives the residuals and reconstructed matrix for a low-rank PCA approximation. An example MATLAB implementation of PCA is available.^[26]
- in GNU Octave, a free software computational environment mostly compatible with MATLAB, the function `princomp`^[27] gives the principal component.
- in the free statistical package R, the functions `princomp`^[28] and `prcomp`^[29] can be used for principal component analysis; `prcomp` uses singular value decomposition which generally gives better numerical accuracy. Recently there has been an explosion in implementations of principal component analysis in various R packages. Some packages that implement PCA in R, include, but are not limited to: `ade4`, `vegan`, `ExPosition`, and `FactoMineR`^[30]
- in SAS, PROC FACTOR offers principal components analysis.
- MLPACK provides an implementation of principal component analysis in C++.
- In *XLMiner*, the Principal Components tab can be used for principal component analysis.^[citation needed]
- In Stata, the `pca` command provides principal components analysis.
- Cornell Spectrum Imager – An open-source toolset built on ImageJ. Enables quick easy PCA analysis for 3D datacubes.^[31]
- imDEV – Free Excel addin to calculate principal components using R package^{[32][33]}
- "ViSta: The Visual Statistics System" – a free software that provides principal components analysis, simple and multiple correspondence analysis.^[34]
- "Spectramap" – software to create a biplot using principal components analysis, correspondence analysis or spectral map analysis.^[35]
- FinMath – a .NET numerical library containing an implementation of PCA.^[36]
- The Unscrambler is a multivariate analysis software enabling Principal Component Analysis (PCA) with PCA Projection.^[citation needed]
- OpenCV^[37]
- NMath, a proprietary numerical library containing PCA for the .NET Framework.^[citation needed]
- In IDL, the principal components can be calculated using the function `pcomp`.^[38]
- Weka computes principal components.^[39]
- Software for analyzing multivariate data with instant response using PCA^[40]
- Orange (software) supports PCA through its Linear Projection widget.^[citation needed]
- A version of PCA adapted for population genetics analysis can be found in the suite EIGENSOFT.^[41]
- PCA can also be performed by the statistical software Partek Genomics Suite.^[42]
- The `libpca` C++ library^[43] offers PCA and corresponding transformations
- Origin contains PCA in its Pro version.

Notes

- [1] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417-441, and 498-520.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **27**, 321-77
- [2] Shaw P.J.A. (2003) *Multivariate statistics for the Environmental Sciences*, Hodder-Arnold. ISBN 0-340-80763-6.
- [3] Jolliffe I.T. Principal Component Analysis ([http://www.springer.com/west/home/new+&+forthcoming+titles+\(default\)?SGWID=4-40356-22-2285433-0](http://www.springer.com/west/home/new+&+forthcoming+titles+(default)?SGWID=4-40356-22-2285433-0)), Series: Springer Series in Statistics (<http://www.springer.com/west/home/statistics/statistical+theory+and+methods?SGWID=4-10129-69-173621571-0>), 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4
- [4] A. A. Miranda, Y. A. Le Borgne, and G. Bontempi. New Routes from Minimal Approximation Error to Principal Components (http://www.ulb.ac.be/di/map/yleborgn/pub/NPL_PCA_07.pdf), Volume 27, Number 3 / June, 2008, Neural Processing Letters, Springer
- [5] Gharib Shirangi, M., History matching production data and uncertainty assessment with a truncated SVD parameterization algorithm, *Journal of Petroleum Science and Engineering*, <http://www.sciencedirect.com/science/article/pii/S0920410513003227>
- [6] Jolliffe, I. T. (2002). *Principal Component Analysis*, second edition Springer-Verlag. ISBN 978-0-387-95442-4.
- [7] Jonathon Shlens, A Tutorial on Principal Component Analysis. (<http://www.sn1.salk.edu/~shlens/pca.pdf>)
- [8] Tech Note
- [9] http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
- [10] A.A. Miranda, Y.-A. Le Borgne, and G. Bontempi. New Routes from Minimal Approximation Error to Principal Components (http://www.ulb.ac.be/di/map/yleborgn/pub/NPL_PCA_07.pdf), Volume 27, Number 3 / June, 2008, Neural Processing Letters, Springer
- [11] eig function (<http://www.mathworks.com/access/helpdesk/help/techdoc/ref/eig.html#998306>) Matlab documentation
- [12] MATLAB PCA-based Face recognition software (<http://www.mathworks.com/matlabcentral/fileexchange/24634>)
- [13] Eigenvalues function (<http://reference.wolfram.com/mathematica/ref/Eigenvalues.html>) Mathematica documentation
- [14] Roweis, Sam. "EM Algorithms for PCA and SPCA." *Advances in Neural Information Processing Systems*. Ed. Michael I. Jordan, Michael J. Kearns, and Sara A. Solla The MIT Press, 1998.
- [15] Kramer,R., (1998) *Chemometric Techniques for Quantitative Analysis* (CRC Press, New York).
- [16] M. Andrecut. Parallel GPU Implementation of Iterative PCA Algorithms. *Journal of Computational Biology*, 16(11), Nov. 2009.
- [17] H. Zha, C. Ding, M. Gu, X. He and H.D. Simon. "Spectral Relaxation for K-means Clustering", <http://ranger.uta.edu/~chqding/papers/Zha-Kmeans.pdf>, *Neural Information Processing Systems vol.14 (NIPS 2001)*. pp. 1057–1064, Vancouver, Canada. Dec. 2001.
- [18] C. Ding and X. He. "K-means Clustering via Principal Component Analysis". *Proc. of Int'l Conf. Machine Learning (ICML 2004)*, pp 225–232. July 2004. <http://ranger.uta.edu/~chqding/papers/KmeansPCA1.pdf>
- [19] <http://www.linkedin.com/groups/What-is-difference-between-factor-107833.S.162765950>
- [20] Timothy A. Brown. *Confirmatory Factor Analysis for Applied Research Methodology in the social sciences*. Guilford Press, 2006
- [21] A. N. Gorban, A. Y. Zinovyev, *Principal Graphs and Manifolds* (<http://arxiv.org/abs/0809.0490>), In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques*, Olivas E.S. et al Eds. Information Science Reference, IGI Global: Hershey, PA, USA, 2009. 28-59.
- [22] Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J. et al.: Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer" *Lancet* 365, 671-679 (2005); Data online (<http://www.ihes.fr/~zinovyev/princmanif2006/>)
- [23] A. Zinovyev, ViDaExpert (<http://bioinfo-out.curie.fr/projects/vidaexpert/>) – Multidimensional Data Visualization Tool (free for non-commercial use). Institut Curie, Paris.
- [24] A.N. Gorban, B. Kegl, D.C. Wunsch, A. Zinovyev (Eds.), *Principal Manifolds for Data Visualisation and Dimension Reduction*, (<http://pca.narod.ru/contentsgkwz.htm>) LNCSE 58, Springer, Berlin – Heidelberg – New York, 2007. ISBN 978-3-540-73749-0
- [25] *PrincipalComponents* (<http://reference.wolfram.com/mathematica/ref/PrincipalComponents.html>) Mathematica Documentation
- [26] *PcaPress* (<http://www.utdallas.edu/~herve/abdi-PCA4Wiley.zip>) www.utdallas.edu
- [27] *princomp* (<http://octave.sourceforge.net/statistics/function/princomp.html>) octave.sourceforge.net
- [28] *princomp* (<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/princomp.html>)
- [29] *prcomp* (<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html>)
- [30] *Multivariate* (<http://cran.r-project.org/web/views/Multivariate.html>) cran.r-project.org
- [31] Cornell Spectrum Imager (<http://code.google.com/p/cornell-spectrum-imager/wiki/Home>)
- [32] *imDEV* (https://sourceforge.net/apps/mediawiki/imdev/index.php?title=Main_Page) sourceforge.net
- [33] *pcaMethods* (<http://www.bioconductor.org/packages/1.9/bioc/html/pcaMethods.html>) www.bioconductor.org
- [34] "ViSta: The Visual Statistics System" (<http://www.mdp.edu.ar/psicologia/vista/vista.htm>) www.mdp.edu.ar
- [35] "Spectramap" (<http://www.coloritto.com>) www.coloritto.com
- [36] *FinMath* (<https://rtmath.net/products/finmath/>) rtmath.net
- [37] *Computer Vision Library* (<http://sourceforge.net/projects/opencvlibrary/>) sourceforge.net
- [38] *PCOMP* (IDL Reference) | Exelis VIS Docs Center (<http://www.exelisvis.com/docs/PCOMP.html>) IDL online documentation
- [39] *javadoc* (<http://weka.sourceforge.net/doc/weka/attributeSelection/PrincipalComponents.html>) weka.sourceforge.net
- [40] Software for analyzing multivariate data with instant response using PCA (<http://www.qlucore.com>) www.qlucore.com

- [41] EIGENSOFT (<http://genepath.med.harvard.edu/~reich/Software.htm>) genepath.med.harvard.edu
- [42] Partek Genomics Suite (<http://www.partek.com/partekgs>) www.partek.com
- [43] <https://sourceforge.net/projects/libpca/>

References

- Jackson, J.E. (1991). *A User's Guide to Principal Components* (Wiley).
- Jolliffe, I. T. (1986). *Principal Component Analysis* ([http://www.springer.com/west/home/new+&+forthcoming+titles+\(default\)?SGWID=4-40356-22-2285433-0](http://www.springer.com/west/home/new+&+forthcoming+titles+(default)?SGWID=4-40356-22-2285433-0)). Springer-Verlag. p. 487. doi: 10.1007/b98835 (<http://dx.doi.org/10.1007/b98835>). ISBN 978-0-387-95442-4.
- Jolliffe, I.T. (2002). *Principal Component Analysis*, second edition (Springer).
- Hao Chen, David L. Reuss, Volker Sick, On the use and interpretation of proper orthogonal decomposition of in-cylinder engine flows. *Measurement Science and Technology*, 2012. 23(8): p. 085302
- Hao Chen, David L. Reuss, David L.S. Hung, Volker Sick, A practical guide for using proper orthogonal decomposition in engine research. *International Journal of Engine Research*, 2013. 14(4): p. 307-319.

External links

- University of Copenhagen video by Rasmus Bro (http://www.youtube.com/watch?v=UUxIXU_Ob6E)
- Stanford University video by Andrew Ng (<http://www.youtube.com/watch?v=ey2PE5xi9-A>)
- A layman's introduction to principal component analysis (<http://www.youtube.com/watch?v=BfTMmoDFXyE>) (a video of less than 100 seconds.)

Diversity index

Diversity index

A **diversity index** is a quantitative measure that reflects how many different types (such as species) there are in a dataset, and simultaneously takes into account how evenly the basic entities (such as individuals) are distributed among those types. The value of a diversity index increases both when the number of types increases and when evenness increases. For a given number of types, the value of a diversity index is maximized when all types are equally abundant.

When diversity indices are used in ecology, the types of interest are usually species, but they can also be other categories, such as genera, families, functional types or haplotypes. The entities of interest are usually individual plants or animals, and the measure of abundance can be, for example, number of individuals, biomass or coverage. In demography, the entities of interest can be people, and the types of interest various demographic groups. In information science, the entities can be characters and the types the different letters of the alphabet. The most commonly used diversity indices are simple transformations of the effective number of types (also known as 'true diversity'), but each diversity index can also be interpreted in its own right as a measure corresponding to some real phenomenon (but a different one for each diversity index).^{[1][2][3][4]}

True diversity (The effective number of types)

True diversity, or the effective number of types, refers to the number of equally-abundant types needed for the average proportional abundance of the types to equal that observed in the dataset of interest (where all types may not be equally abundant). The true diversity in a dataset is calculated by first taking the weighted generalized mean of the proportional abundances of the types in the dataset, and then taking the inverse of this. The equation is:

$${}^qD = \frac{1}{\sqrt[q-1]{\sum_{i=1}^R p_i p_i^{q-1}}}$$

The denominator equals average proportional abundance of the types in the dataset as calculated with the weighted generalized mean with exponent $q - 1$. In the equation, R is richness (the total number of types in the dataset), and the proportional abundance of the i th type is p_i . The proportional abundances themselves are used as the nominal weights. When $q = 1$, the above equation is undefined, so the corresponding mean is calculated with the following equation instead:

$${}^1D = \frac{1}{\prod_{i=1}^R p_i^{p_i}}$$

The value of q is often referred to as the order of the diversity. It defines the sensitivity of the diversity value to rare vs. abundant species by modifying how the mean of the species proportional abundances is calculated. With some values of the parameter q , the generalized mean with exponent $q - 1$ gives familiar kinds of mean as special cases. In particular, $q = 0$ corresponds to the harmonic mean, $q = 1$ to the geometric mean and $q = 2$ to the arithmetic mean. As q approaches infinity, the generalized mean with exponent $q - 1$ approaches the maximum p_i value, which is the proportional abundance of the most abundant species in the dataset. In practice, increasing the value of q hence increases the effective weight given to the most abundant species. This leads to obtaining a larger mean p_i value and a smaller true diversity (qD) value.

When $q = 1$, the geometric mean of the p_i values is used, and each species is exactly weighted by its proportional abundance (in the geometric mean, weights are the exponents). When $q > 1$, the weight given to abundant species is

exaggerated, and when $q < 1$, the weight given to rare species is. At $q = 0$, the species weights exactly cancel out the species proportional abundances, such that mean p_i equals $1 / R$ even when all species are not equally abundant. At $q = 0$, the effective number of species, 0D , hence equals the actual number of species (R). In the context of diversity, q is generally limited to non-negative values. This is because negative values of q would give rare species so much more weight than abundant ones that qD would exceed R .

The general equation of diversity is often written in the form:

$${}^qD = \left(\sum_{i=1}^R p_i^q \right)^{1/(1-q)}$$

The term inside the parentheses is called the basic sum. Some popular diversity indices correspond to the basic sum as calculated with different values of q .

For diversity of order one, an alternative equation is:

$${}^1D = \exp \left(- \sum_{i=1}^R p_i \ln p_i \right) = \exp(H')$$

where H' is the Shannon index as calculated with natural logarithms (see below).

Richness

Richness R simply quantifies how many different types the dataset of interest contains. For example, species richness (usually notated S) of a dataset is the number of different species in the corresponding species list. Richness is a simple measure, so it has been a popular diversity index in ecology, where abundance data are often not available for the datasets of interest. Because richness does not take the abundances of the types into account, it is not the same thing as diversity, which does take abundances into account. However, if true diversity is calculated with $q = 0$, the effective number of types (0D) equals the actual number of types (R).

Shannon index

The Shannon index has been a popular diversity index in the ecological literature, where it is also known as Shannon's diversity index, the Shannon–Wiener index, the Shannon–Weaver index and the Shannon entropy. The measure was originally proposed by Claude Shannon to quantify the entropy (uncertainty or information content) in strings of text.^[5] The idea is that the more different letters there are, and the more equal their proportional abundances in the string of interest, the more difficult it is to correctly predict which letter will be the next one in the string. The Shannon entropy quantifies the uncertainty (entropy or degree of surprise) associated with this prediction. It is most often calculated as follows:

$$H' = - \sum_{i=1}^R p_i \ln p_i$$

where p_i is the proportion of characters belonging to the i th type of letter in the string of interest. In ecology, p_i is often the proportion of individuals belonging to the i th species in the dataset of interest. Then the Shannon entropy quantifies the uncertainty in predicting the species identity of an individual that is taken at random from the dataset.

Although the equation is here written with natural logarithms, the base of the logarithm used when calculating the Shannon entropy can be chosen freely. Shannon himself discussed logarithm bases 2, 10 and e , and these have since become the most popular bases in applications that use the Shannon entropy. Each log base corresponds to a different measurement unit, which have been called binary digits (bits), decimal digits (decits) and natural digits (nats) for the bases 2, 10 and e , respectively. Comparing Shannon entropy values that were originally calculated with different log bases requires converting them to the same log base: change from the base a to base b is obtained with multiplication by $\log_b a$.

It has been shown that the Shannon index is based on the weighted geometric mean of the proportional abundances of the types, and that it equals the logarithm of true diversity as calculated with $q = 1$:

$$H' = - \sum_{i=1}^R p_i \ln p_i = - \sum_{i=1}^R \ln p_i^{p_i}$$

This can also be written

$$H' = -(\ln p_1^{p_1} + \ln p_2^{p_2} + \ln p_3^{p_3} + \dots + \ln p_R^{p_R})$$

which equals

$$H' = - \ln p_1^{p_1} p_2^{p_2} p_3^{p_3} \dots p_R^{p_R} = \ln \left(\frac{1}{p_1^{p_1} p_2^{p_2} p_3^{p_3} \dots p_R^{p_R}} \right) = \ln \left(\frac{1}{\prod_{i=1}^R p_i^{p_i}} \right)$$

Since the sum of the p_i values equals unity by definition, the denominator equals the weighted geometric mean of the p_i values, with the p_i values themselves being used as the weights (exponents in the equation). The term within the parentheses hence equals true diversity 1D , and H' equals $\ln({}^1D)$.

When all types in the dataset of interest are equally common, all p_i values equal $1/R$, and the Shannon index hence takes the value $\ln(R)$. The more unequal the abundances of the types, the larger the weighted geometric mean of the p_i values, and the smaller the corresponding Shannon entropy. If practically all abundance is concentrated to one type, and the other types are very rare (even if there are many of them), Shannon entropy approaches zero. When there is only one type in the dataset, Shannon entropy exactly equals zero (there is no uncertainty in predicting the type of the next randomly chosen entity).

Simpson index

The Simpson index was introduced in 1949 by Edward H. Simpson to measure the degree of concentration when individuals are classified into types.^[6] The same index was rediscovered by Orris C. Herfindahl in 1950.^[7] The square root of the index had already been introduced in 1945 by the economist Albert O. Hirschman.^[8] As a result, the same measure is usually known as the Simpson index in ecology, and as the Herfindahl index or the Herfindahl–Hirschman index (HHI) in economics.

The measure equals the probability that two entities taken at random from the dataset of interest represent the same type. It equals:

$$\lambda = \sum_{i=1}^R p_i^2$$

This also equals the weighted arithmetic mean of the proportional abundances p_i of the types of interest, with the proportional abundances themselves being used as the weights. Proportional abundances are by definition constrained to values between zero and unity, but their weighted arithmetic mean, and hence λ , can never be smaller than $1/S$, which is reached when all types are equally abundant.

By comparing the equation used to calculate λ with the equations used to calculate true diversity, it can be seen that $1/\lambda$ equals 2D , i.e. true diversity as calculated with $q = 2$. The original Simpson's index hence equals the corresponding basic sum.

The interpretation of λ as the probability that two entities taken at random from the dataset of interest represent the same type assumes that the first entity is replaced to the dataset before taking the second entity. If the dataset is very large, sampling without replacement gives approximately the same result, but in small datasets the difference can be substantial. If the dataset is small, and sampling without replacement is assumed, the probability of obtaining the same type with both random draws is:

$$l = \frac{\sum_{i=1}^R n_i(n_i - 1)}{N(N - 1)}$$

where n_i is the number of entities belonging to the i th type and N is the total number of entities in the dataset. This form of the Simpson index is also known as the Hunter–Gaston index in microbiology.^[9]

Since mean proportional abundance of the types increases with decreasing number of types and increasing abundance of the most abundant type, λ obtains small values in datasets of high diversity and large values in datasets of low diversity. This is counterintuitive behavior for a diversity index, so often such transformations of λ that increase with increasing diversity have been used instead. The most popular of such indices have been the inverse Simpson index ($1/\lambda$) and the Gini–Simpson index ($1 - \lambda$). Both of these have also been called the Simpson index in the ecological literature, so care is needed to avoid accidentally comparing the different indices as if they were the same.

Inverse Simpson index

The inverse Simpson index equals:

$$1/\lambda = \frac{1}{\sum_{i=1}^R p_i^2} = {}^2D$$

This simply equals true diversity of order 2, i.e. the effective number of types that is obtained when the weighted arithmetic mean is used to quantify average proportional abundance of types in the dataset of interest.

Gini–Simpson index

The original Simpson index λ equals the probability that two entities taken at random from the dataset of interest (with replacement) represent the same type. Its transformation $1 - \lambda$ therefore equals the probability that the two entities represent different types. This measure is also known in ecology as the probability of interspecific encounter (*PIE*)^[10] and the Gini–Simpson index. It can be expressed as a transformation of true diversity of order 2:

$$1 - \lambda = 1 - \sum_{i=1}^R p_i^2 = 1 - 1/{}^2D$$

The Gibbs–Martin index of sociology, psychology and management studies,^[11] which is also known as the Blau index, is the same measure as the Gini–Simpson index.

Berger–Parker index

The Berger–Parker index equals the maximum p_i value in the dataset, i.e. the proportional abundance of the most abundant type. This corresponds to the weighted generalized mean of the p_i values when q approaches infinity, and hence equals the inverse of true diversity of order infinity ($1/{}^\infty D$).

Rényi entropy

The Rényi entropy is a generalization of the Shannon entropy to other values of q than unity. It can be expressed:

$${}^q H = \frac{1}{1 - q} \ln \left(\sum_{i=1}^R p_i^q \right)$$

which equals

$${}^q H = \ln \left(\frac{1}{\sqrt[q-1]{\sum_{i=1}^R p_i p_i^{q-1}}} \right) = \ln({}^q D)$$

This means that taking the logarithm of true diversity based on any value of q gives the Rényi entropy corresponding to the same value of q .

References

- [1] Hill, M. O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54, 427–432. (<http://dx.doi.org/10.2307/1934352>)
- [2] Jost, L. (2006) Entropy and diversity. *Oikos*, 113, 363–375.
- [3] Tuomisto, H. (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*, 33, 2–22.
- [4] Tuomisto, H. 2010. "A consistent terminology for quantifying species diversity? Yes, it does exist". *Oecologia* 4: 853–860.
- [5] Shannon, C. E. (1948) A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423 and 623–656.
- [6] Simpson, E. H. (1949) Measurement of diversity. *Nature*, 163, 688.
- [7] Herfindahl, O. C. (1950) Concentration in the U.S. Steel Industry. Unpublished doctoral dissertation, Columbia University.
- [8] Hirschman, A. O. (1945) National power and the structure of foreign trade. Berkeley.
- [9] Hunter PR, Gaston MA (1988) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J Clin Microbiol* 26(11): 2465–2466
- [10] Hurlbert, S.H. (1971) The nonconcept of species diversity: A critique and alternative parameters. *Ecology* 52, 577–586. (<http://dx.doi.org/10.2307/1934145>)
- [11] Gibbs, Jack P., and William T. Martin, 1962. Urbanization, technology and the division of labor. *American Sociological Review* 27: 667–677.

Further reading

- Colinvaux, Paul A. (1973). *Introduction to Ecology*. Wiley. ISBN 0-471-16498-4.
- Cover, Thomas M.; and Thomas, Joy A. (1991). *Elements of Information Theory*. Wiley. ISBN 0-471-06259-6. See *chapter 5* for an elaboration of coding procedures described informally above.
- Chao, A.; Shen, T-J. (2003) "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample" (http://chao.stat.nthu.edu.tw/paper/2003_EEST_10_P429.pdf), *Environmental and Ecological Statistics*, 10 (4),429–443 doi: 10.1023/A:1026096204727 (<http://dx.doi.org/10.1023/A:1026096204727>)

External links

- Simpson's Diversity index (<http://www.countrysideinfo.co.uk/simpsons.htm>)
- Diversity indices (<http://www.tiem.utk.edu/~gross/bioed/bealsmodules/simpsonDI.html>) gives some examples of estimates of Simpson's index for real ecosystems.

Clustering

Hierarchical clustering

In data mining, **hierarchical clustering** is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: ^[citation needed]

- **Agglomerative:** This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive:** This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.

In the general case, the complexity of agglomerative clustering is $\mathcal{O}(n^3)$, which makes them too slow for large data sets. Divisive clustering with an exhaustive search is $\mathcal{O}(2^n)$, which is even worse. However, for some special cases, optimal efficient agglomerative methods (of complexity $\mathcal{O}(n^2)$) are known: SLINK for single-linkage and CLINK for complete-linkage clustering.

Cluster dissimilarity

In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

Metric

The choice of an appropriate metric will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. For example, in a 2-dimensional space, the distance between the point (1,0) and the origin (0,0) is always 1 according to the usual norms, but the distance between the point (1,1) and the origin (0,0) can be 2, $\sqrt{2}$ or 1 under Manhattan distance, Euclidean distance or maximum distance respectively.

Some commonly used metrics for hierarchical clustering are:

Names	Formula
Euclidean distance	$\ a - b\ _2 = \sqrt{\sum_i (a_i - b_i)^2}$
squared Euclidean distance	$\ a - b\ _2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\ a - b\ _1 = \sum_i a_i - b_i $
maximum distance	$\ a - b\ _\infty = \max_i a_i - b_i $
Mahalanobis distance	$\sqrt{(a - b)^T S^{-1} (a - b)}$ where S is the covariance matrix
cosine similarity	$\frac{a \cdot b}{\ a\ \ b\ }$

For text or other non-numeric data, metrics such as the Hamming distance or Levenshtein distance are often used.

A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance.^[citation needed]

Linkage criteria

The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations.

Some commonly used linkage criteria between two sets of observations A and B are:^[1]

Names	Formula
Maximum or complete linkage clustering	$\max \{ d(a, b) : a \in A, b \in B \}.$
Minimum or single-linkage clustering	$\min \{ d(a, b) : a \in A, b \in B \}.$
Mean or average linkage clustering, or UPGMA	$\frac{1}{ A B } \sum_{a \in A} \sum_{b \in B} d(a, b).$
Minimum energy clustering	$\frac{2}{nm} \sum_{i,j=1}^{n,m} \ a_i - b_j\ _2 - \frac{1}{n^2} \sum_{i,j=1}^n \ a_i - a_j\ _2 - \frac{1}{m^2} \sum_{i,j=1}^m \ b_i - b_j\ _2$

where d is the chosen metric. Other linkage criteria include:

- The sum of all intra-cluster variance.
- The decrease in variance for the cluster being merged (Ward's criterion).
- The probability that candidate clusters spawn from the same distribution function (V-linkage).
- The product of in-degree and out-degree on a k-nearest-neighbor graph (graph degree linkage).^[2]
- The increment of some cluster descriptor (i.e., a quantity defined for measuring the quality of a cluster) after merging two clusters.^{[3] [4] [5]}

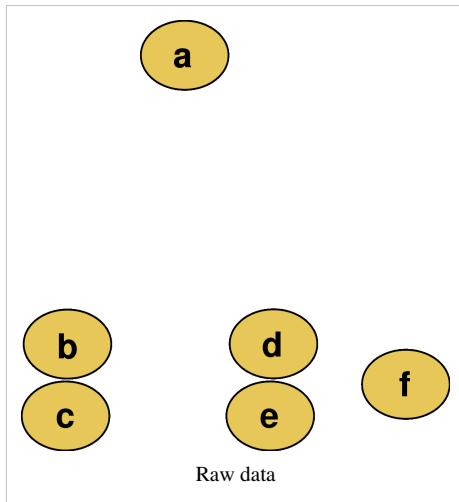
Discussion

Hierarchical clustering has the distinct advantage that any valid measure of distance can be used. In fact, the observations themselves are not required: all that is used is a matrix of distances.

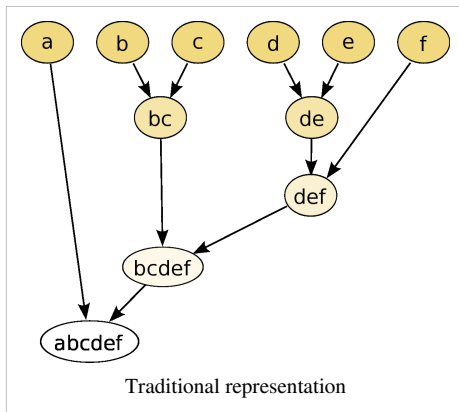
Example for Agglomerative Clustering

For example, suppose this data is to be clustered, and the Euclidean distance is the distance metric.

Cutting the tree at a given height will give a partitioning clustering at a selected precision. In this example, cutting after the second row of the dendrogram will yield clusters $\{a\}$ $\{b\ c\}$ $\{d\ e\}$ $\{f\}$. Cutting after the third row will yield clusters $\{a\}$ $\{b\ c\}$ $\{d\ e\ f\}$, which is a coarser clustering, with a smaller number of larger clusters.



The hierarchical clustering dendrogram would be as such:



This method builds the hierarchy from the individual elements by progressively merging clusters. In our example, we have six elements $\{a\}$ $\{b\}$ $\{c\}$ $\{d\}$ $\{e\}$ and $\{f\}$. The first step is to determine which elements to merge in a cluster. Usually, we want to take the two closest elements, according to the chosen distance.

Optionally, one can also construct a distance matrix at this stage, where the number in the i -th row j -th column is the distance between the i -th and j -th elements. Then, as clustering progresses, rows and columns are merged as the clusters are merged and the distances updated. This is a common way to implement this type of clustering, and has the benefit of caching distances between clusters. A simple agglomerative clustering algorithm is described in the single-linkage clustering page; it can easily be adapted to different types of linkage (see below).

Suppose we have merged the two closest elements b and c , we now have the following clusters $\{a\}$, $\{b, c\}$, $\{d\}$, $\{e\}$ and $\{f\}$, and want to merge them further. To do that, we need to take the distance between $\{a\}$ and $\{b\ c\}$, and therefore define the distance between two clusters. Usually the distance between two clusters \mathcal{A} and \mathcal{B} is one of

the following:

- The maximum distance between elements of each cluster (also called complete-linkage clustering):

$$\max\{d(x, y) : x \in \mathcal{A}, y \in \mathcal{B}\}.$$

- The minimum distance between elements of each cluster (also called single-linkage clustering):

$$\min\{d(x, y) : x \in \mathcal{A}, y \in \mathcal{B}\}.$$

- The mean distance between elements of each cluster (also called average linkage clustering, used e.g. in UPGMA):

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y).$$

- The sum of all intra-cluster variance.
- The increase in variance for the cluster being merged (Ward's method^[6])
- The probability that candidate clusters spawn from the same distribution function (V-linkage).

Each agglomeration occurs at a greater distance between clusters than the previous agglomeration, and one can decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion).

Software

Open Source Frameworks

- Cluster 3.0^[6] provides a nice Graphical User Interface to access to different clustering routines and is available for Windows, Mac OS X, Linux, Unix.
- ELKI includes multiple hierarchical clustering algorithms, various linkage strategies and also includes the efficient SLINK algorithm, flexible cluster extraction from dendrograms and various other cluster analysis algorithms.
- Octave, the GNU analog to MATLAB implements hierarchical clustering in linkage function^[7]
- Orange, a free data mining software suite, module orngClustering^[8] for scripting in Python, or cluster analysis through visual programming.
- R has several functions for hierarchical clustering: see CRAN Task View: Cluster Analysis & Finite Mixture Models^[9] for more information.
- scikit-learn implements a hierarchical clustering based on the Ward algorithm only.
- Weka includes hierarchical cluster analysis.

Standalone implementations

- CrimeStat implements two hierarchical clustering routines, a nearest neighbor (Nnh) and a risk-adjusted(Rnnh).
- figure^[10] is a JavaScript package that implements some agglomerative clustering functions (single-linkage, complete-linkage, average-linkage) and functions to visualize clustering output (e.g. dendrograms).
- hcluster^[11] is a Python implementation, based on NumPy, which supports hierarchical clustering and plotting.
- Hierarchical Agglomerative Clustering^[12] implemented as C# visual studio project that includes real text files processing, building of document-term matrix with stop words filtering and stemming.
- MultiDendrograms^[13] An open source Java application for variable-group agglomerative hierarchical clustering, with graphical user interface.
- Graph Agglomerative Clustering (GAC) toolbox^[14] implemented several graph-based agglomerative clustering algorithms.

Commercial

- MATLAB includes hierarchical cluster analysis.
- SAS includes hierarchical cluster analysis.
- Mathematica includes a Hierarchical Clustering Package

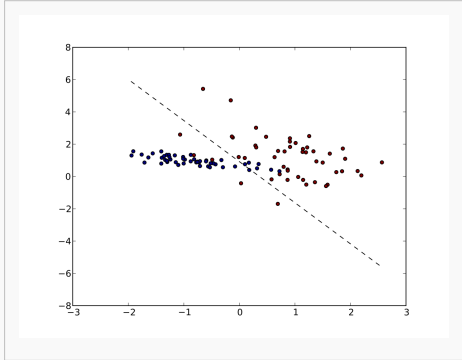
Notes



- [1] Székely, G. J. and Rizzo, M. L. (2005) Hierarchical clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method, *Journal of Classification* 22, 151-183.
- [2] Zhang, et al. "Graph degree linkage: Agglomerative clustering on a directed graph." 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012. <http://arxiv.org/abs/1208.5092>
- [3] Zhang, et al. "Agglomerative clustering via maximum incremental path integral." *Pattern Recognition* (2013).
- [4] Zhao, and Tang. "Cyclizing clusters via zeta function of a graph." *Advances in Neural Information Processing Systems*. 2008.
- [5] Ma, et al. "Segmentation of multivariate mixed data via lossy data coding and compression." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9) (2007): 1546-1562.
- [6] <http://bonsai.hgc.jp/~mdehoon/software/cluster/>
- [7] <http://octave.sourceforge.net/statistics/function/linkage.html>
- [8] <http://www.ailab.si/orange/doc/modules/orngClustering.htm>
- [9] <http://cran.r-project.org/web/views/Cluster.html>
- [10] <http://code.google.com/p/figue/>
- [11] <http://code.google.com/p/scipy-cluster/>
- [12] <http://www.semanticsearchart.com/researchHAC.html>
- [13] <http://deim.urv.cat/~sgomez/multidendrograms.php>
- [14] <http://www.mathworks.com/matlabcentral/fileexchange/38018-graph-agglomerative-clustering-gac-toolbox>

References and further reading

- Kaufman, L.; Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis* (1 ed.). New York: John Wiley. ISBN 0-471-87876-6.
- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "14.3.12 Hierarchical clustering" (<http://www-stat.stanford.edu/~tibs/ElemStatLearn/>) (PDF). *The Elements of Statistical Learning* (2nd ed.). New York: Springer. pp. 520–528. ISBN 0-387-84857-6. Retrieved 2009-10-20.
- Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP (2007). "Section 16.4. Hierarchical Clustering by Phylogenetic Trees" (<http://apps.nrbook.com/empanel/index.html#pg=868>). *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8.

K-means clustering

Machine learning and data mining	
	
Problems	
<ul style="list-style-type: none"> • • • • • • • • • • • • • • 	<ul style="list-style-type: none"> Classification Clustering Regression Anomaly detection Association rules Reinforcement learning Structured prediction Feature learning Online learning Semi-supervised learning Grammar induction
Supervised learning (classification • regression)	
<ul style="list-style-type: none"> • • • • • • • • • • 	<ul style="list-style-type: none"> Decision trees Ensembles (Bagging, Boosting, Random forest) k-NN Linear regression Naive Bayes Neural networks Logistic regression Perceptron Support vector machine (SVM)
Clustering	
<ul style="list-style-type: none"> • • • • • • • 	<ul style="list-style-type: none"> BIRCH Hierarchical k-means Expectation-maximization (EM) DBSCAN OPTICS Mean-shift
Dimensionality reduction	

<ul style="list-style-type: none"> • Factor analysis • CCA • ICA • LDA • NMF • PCA 	
Structured prediction	
<ul style="list-style-type: none"> • Graphical models (CRF, HMM) 	
Anomaly detection	
<ul style="list-style-type: none"> • k-NN • Local outlier factor 	
Theory	
<ul style="list-style-type: none"> • Bias-variance dilemma • Computational learning theory • Empirical risk minimization • PAC learning • VC theory 	
<ul style="list-style-type: none"> •  Computer science portal •  Statistics portal 	
<ul style="list-style-type: none"> • • • 	v t $e^{[1]}$

k -means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k -means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, k -means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

Description

Given a set of observations $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, where each observation is a d -dimensional real vector, k -means clustering aims to partition the n observations into k sets ($k \leq n$) $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i .

History

The term "*k*-means" was first used by James MacQueen in 1967, though the idea goes back to Hugo Steinhaus in 1957. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published outside of Bell Labs until 1982.^[2] In 1965, E.W.Forgy published essentially the same method, which is why it is sometimes referred to as Lloyd-Forgy. A more efficient version was proposed and published in Fortran by Hartigan and Wong in 1975/1979.

Algorithms

Standard algorithm

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the ***k*-means algorithm**; it is also referred to as **Lloyd's algorithm**, particularly in the computer science community.

Given an initial set of *k* means $m_1^{(1)}, \dots, m_k^{(1)}$ (see below), the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean.^[3] (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \forall j, 1 \leq j \leq k\},$$

where each x_p is assigned to exactly one $S_i^{(t)}$, even if it could be assigned to two or more of them.

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

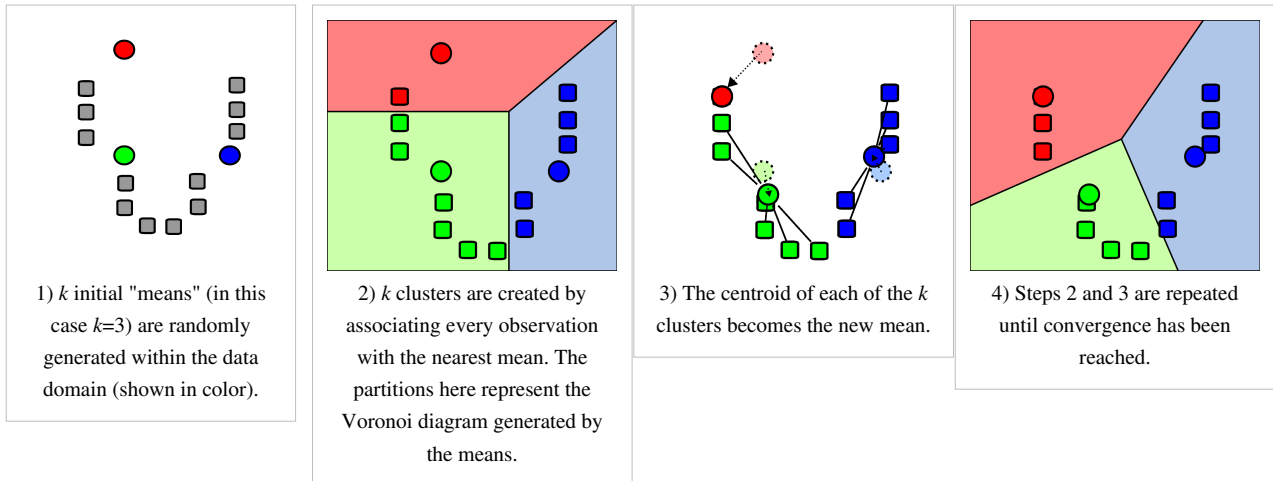
The algorithm has converged when the assignments no longer change. Since both steps optimize the WCSS objective, and there only exists a finite number of such partitionings, the algorithm must converge to a (local) optimum. There is no guarantee that the global optimum is found using this algorithm.

The algorithm is often presented as assigning objects to the nearest cluster by distance. This is slightly inaccurate: the algorithm aims at minimizing the WCSS objective, and thus assigns by "least sum of squares". Using a different distance function other than (squared) Euclidean distance may stop the algorithm from converging. It is correct that the smallest Euclidean distance yields the smallest squared Euclidean distance and thus also yields the smallest sum of squares. Various modifications of *k*-means such as spherical *k*-means and *k*-medoids have been proposed to allow using other distance measures.

Initialization methods

Commonly used initialization methods are Forgy and Random Partition. The Forgy method randomly chooses *k* observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster's randomly assigned points. The Forgy method tends to spread the initial means out, while Random Partition places all of them close to the center of the data set. According to Hamerly et al., the Random Partition method is generally preferable for algorithms such as the *k*-harmonic means and fuzzy *k*-means. For expectation maximization and standard *k*-means algorithms, the Forgy method of initialization is preferable.

Demonstration of the standard algorithm



As it is a heuristic algorithm, there is no guarantee that it will converge to the global optimum, and the result may depend on the initial clusters. As the algorithm is usually very fast, it is common to run it multiple times with different starting conditions. However, in the worst case, k -means can be very slow to converge: in particular it has been shown that there exist certain point sets, even in 2 dimensions, on which k -means takes exponential time, that is $2^{\Omega(n)}$, to converge. These point sets do not seem to arise in practice: this is corroborated by the fact that the smoothed running time of k -means is polynomial.

The "assignment" step is also referred to as **expectation step**, the "update step" as **maximization step**, making this algorithm a variant of the *generalized* expectation-maximization algorithm.

Complexity

Regarding computational complexity, finding the optimal solution to the k -means clustering problem for observations in d dimensions is:

- NP-hard in general Euclidean space d even for 2 clusters
- NP-hard for a general number of clusters k even in the plane
- If k and d (the dimension) are fixed, the problem can be exactly solved in time $O(n^{dk+1} \log n)$, where n is the number of entities to be clustered

Thus, a variety of heuristic algorithms such as Lloyds algorithm given above are generally used.

- Lloyd's k -means algorithm has polynomial smoothed running time. It is shown that for arbitrary set of n points in $[0, 1]^d$, if each point is independently perturbed by a normal distribution with mean 0 and variance σ^2 , then the expected running time of k -means algorithm is bounded by $O(n^{3d} k^{3d} d^8 \log^4(n) / \sigma^6)$, which is a polynomial in n, k, d and $1/\sigma$.
- Better bounds are proved for simple cases. For example, showed that the running time of k -means algorithm is bounded by $O(dn^4 M^2)$ for n points in an integer lattice $\{1, \dots, M\}^d$.

Variations

- k-medians clustering uses the median in each dimension instead of the mean, and this way minimizes L_1 norm (Taxicab geometry).
- k-medoids (also: Partitioning Around Medoids, PAM) uses the medoid instead of the mean, and this way minimizes the sum of distances for *arbitrary* distance functions.
- Fuzzy C-Means Clustering is a soft version of K-means, where each data point has a fuzzy degree of belonging to each cluster.
- Gaussian mixture models trained with expectation-maximization algorithm (EM algorithm) maintains probabilistic assignments to clusters, instead of deterministic assignments, and multivariate Gaussian distributions instead of means.
- Several methods have been proposed to choose better starting clusters. One recent proposal is k-means++.
- The filtering algorithm uses kd-trees to speed up each k-means step.
- Some methods attempt to speed up each k-means step using coresets or the triangle inequality.
- Escape local optima by swapping points between clusters.
- The Spherical k-means clustering algorithm is suitable for directional data.
- The Minkowski metric weighted k-means deals with irrelevant features by assigning cluster specific weights to each feature

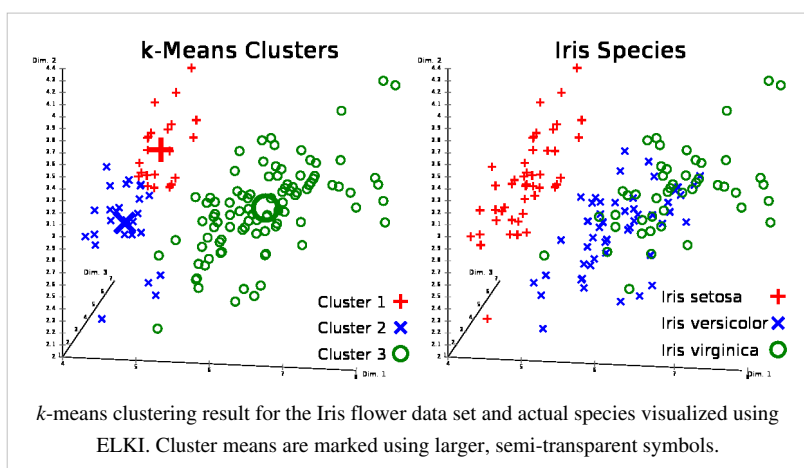
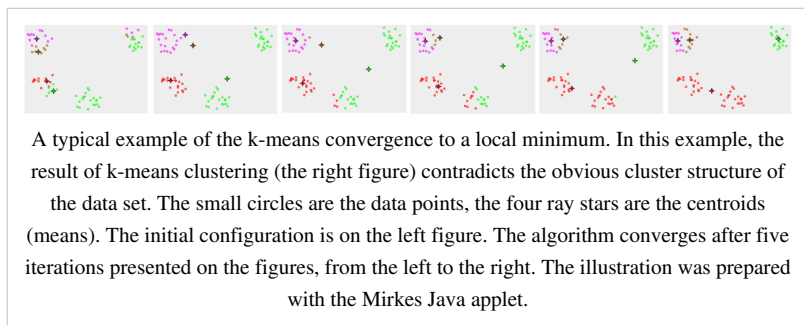
Discussion

The two key features of k -means which make it efficient are often regarded as its biggest drawbacks:

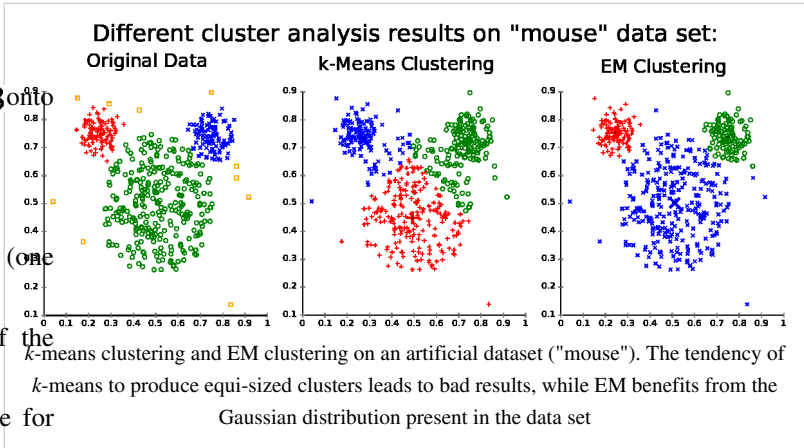
- Euclidean distance is used as a metric and variance is used as a measure of cluster scatter.
- The number of clusters k is an input parameter: an inappropriate choice of k may yield poor results. That is why, when performing k -means, it is important to run diagnostic checks for determining the number of clusters in the data set.
- Convergence to a local minimum may produce counterintuitive ("wrong") results (see example in Fig.).

A key limitation of k -means is its cluster model. The concept is based on spherical clusters that are separable in a way so that the mean value

converges towards the cluster center. The clusters are expected to be of similar size, so that the assignment to



the nearest cluster center is the correct assignment. When for example applying k -means with a value of $k = 3$ on the well-known Iris flower data set, the result often fails to separate the three Iris species contained in the data set. With $k = 2$, the two visible clusters (one containing two species) will be discovered, whereas with $k = 3$ one of the two clusters will be split into two even parts. In fact, $k = 2$ is more appropriate for this data set, despite the data set



k-means clustering and EM clustering on an artificial dataset ("mouse"). The tendency of *k*-means to produce equi-sized clusters leads to bad results, while EM benefits from the Gaussian distribution present in the data set

containing 3 classes. As with any other clustering algorithm, the *k*-means result relies on the data set to satisfy the assumptions made by the clustering algorithms. It works well on some data sets, while failing on others.

The result of *k*-means can also be seen as the Voronoi cells of the cluster means. Since data is split halfway between cluster means, this can lead to suboptimal splits as can be seen in the "mouse" example. The Gaussian models used by the Expectation-maximization algorithm (which can be seen as a generalization of *k*-means) are more flexible here by having both variances and covariances. The EM result is thus able to accommodate clusters of variable size much better than *k*-means as well as correlated clusters (not in this example).

Applications

k-means clustering in particular when using heuristics such as Lloyd's algorithm is rather easy to implement and apply even on large data sets. As such, it has been successfully used in various topics, ranging from market segmentation, computer vision, geostatistics,^[4] and astronomy to agriculture. It often is used as a preprocessing step for other algorithms, for example to find a starting configuration.

Vector quantization

k-means originates from signal processing, and still finds use in this domain. For example in computer graphics, color quantization is the task of reducing the color palette of an image to a fixed number of colors k . The *k*-means algorithm can easily be used for this task and produces competitive results. Other uses of vector quantization include non-random sampling, as *k*-means can easily be used to choose k different but prototypical objects from a large data set for further analysis.

Cluster analysis

In cluster analysis, the *k*-means algorithm can be used to partition the input data set into k partitions (clusters).

However, the pure *k*-means algorithm is not very flexible, and as such of limited use (except



Two-channel (for illustration purposes -- red and green only) color image.

for when

vector quantization as above is actually the desired use case!). In particular, the parameter k is known to be hard to choose (as discussed below) when not given by external constraints. In contrast to other algorithms, k -means can also not be used with arbitrary distance functions or be use on non-numerical data. For these use cases, many other algorithms have been developed since.

Feature learning

k -means clustering has been used as a feature learning (or dictionary learning) step, which can be used in the for (semi-)supervised learning or unsupervised learning. The basic approach is first to train a k -means clustering representation, using the input training data (which need not be labelled). Then, to project any input datum into the new feature space, we have a choice of "encoding" functions, but we can use for example the

thresholded matrix-product of the datum with the centroid locations, the distance from the datum to each centroid, or simply an indicator function for the nearest centroid, or some smooth transformation of the distance. Alternatively, by transforming the sample-cluster distance through a Gaussian RBF, one effectively obtains the hidden layer of a radial basis function network.

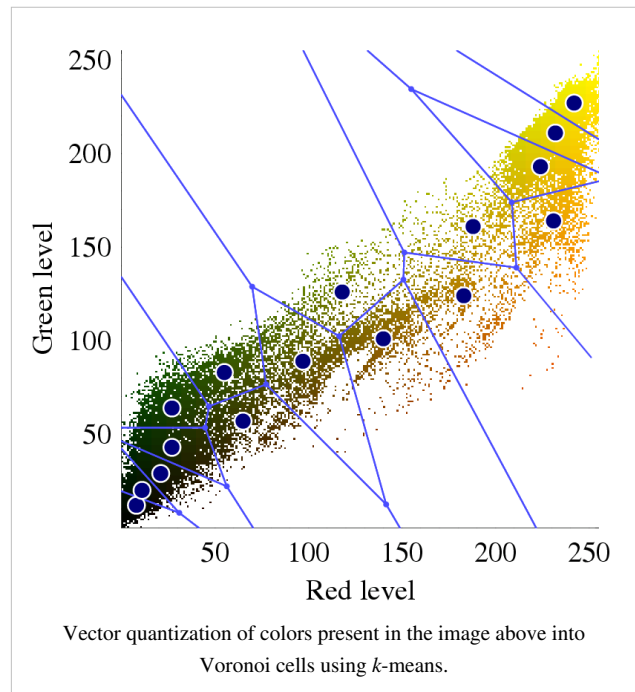
This use of k -means has been successfully combined with simple, linear classifiers for semi-supervised learning in NLP (specifically for named entity recognition) and in computer vision. On an object recognition task, it was found to exhibit comparable performance with more sophisticated feature learning approaches such as autoencoders and restricted Boltzmann machines. However, it generally requires more data than the sophisticated methods, for equivalent performance, because each data point only contributes to one "feature" rather than multiple.

Relation to other statistical machine learning algorithms

k -means clustering, and its associated expectation-maximization algorithm, is a special case of a Gaussian mixture model, specifically, the limit of taking all covariances as diagonal, equal, and small. It is often easy to generalize a k -means problem into a Gaussian mixture model. Another generalization of the k -means algorithm is the K-SVD algorithm, which estimates data points as a sparse linear combination of "codebook vectors". K-means corresponds to the special case of using a single codebook vector, with a weight of 1.

Mean shift clustering

Basic mean shift clustering algorithms maintain a set of data points the same size as the input data set. Initially, this set is copied from the input set. Then this set is iteratively replaced by the mean of those points in the set that are within a given distance of that point. By contrast, k -means restricts this updated set to k points usually much less than the number of points in the input data set, and replaces each point in this set by the mean of all points in the *input set* that are closer to that point than any other (e.g. within the Voronoi partition of each updating point). A mean shift algorithm that is similar then to k -means, called *likelihood mean shift*, replaces the set of points undergoing replacement by the mean of all points in the input set that are within a given distance of the changing set. One of the advantages of mean shift over k -means is that there is no need to choose the number of clusters, because mean shift is likely to find only a few clusters if indeed only a small number exist. However, mean shift can be much slower



than k -means, and still requires selection of a bandwidth parameter. Mean shift has soft variants much as k -means does.

Principal component analysis (PCA)

It was asserted in that the relaxed solution of k -means clustering, specified by the cluster indicators, is given by the PCA (principal component analysis) principal components, and the PCA subspace spanned by the principal directions is identical to the cluster centroid subspace. However, that PCA is a useful relaxation of k -means clustering was not a new result (see, for example,), and it is straightforward to uncover counterexamples to the statement that the cluster centroid subspace is spanned by the principal directions^[citation needed].

Bilateral filtering

k -means implicitly assumes that the ordering of the input data set does not matter. The bilateral filter is similar to K -means and mean shift in that it maintains a set of data points that are iteratively replaced by means. However, the bilateral filter restricts the calculation of the (kernel weighted) mean to include only points that are close in the ordering of the input data. This makes it applicable to problems such as image denoising, where the spatial arrangement of pixels in an image is of critical importance.

Similar problems

The set of squared error minimizing cluster functions also includes the k -medoids algorithm, an approach which forces the center point of each cluster to be one of the actual points, i.e., it uses medoids in place of centroids.

Software

Free

- Apache Mahout k -Means^[5]
- CrimeStat implements two spatial K -means algorithms, one of which allows the user to define the starting locations.
- ELKI contains k -means (with Lloyd and MacQueen iteration, along with different initializations such as k -means++ initialization) and various more advanced clustering algorithms
- MLPACK contains a C++ implementation of k -means
- R k means^[6] implements a variety of algorithms
- SciPy vector-quantization^[7]
- Scikit-learn implements a popular python machine-learning library which contains various clustering algorithms
- Silverlight widget demonstrating k -means algorithm^[8]
- PostgreSQL extension for k -means^[9]
- CMU's GraphLab Clustering library^[10] Efficient multicore implementation for large scale data.
- Weka contains k -means and a few variants of it, including k -means++ and x -means.
- Spectral Python^[11] contains methods for unsupervised classification including a K -means clustering method.
- scikit learn^[12] machine learning in Python contains a K -Means implementation
- OpenCV contains a K -means^[13] implementation under BSD licence.
- Yael^[14] includes an efficient multi-threaded C implementation of k -means, with C, Python and Matlab interfaces.

Commercial

- IDL Cluster, Clust_Wts
- *Mathematica* ClusteringComponents function ^[15]
- MATLAB kmeans ^[16]
- SAS FASTCLUS ^[17]
- Stata kmeans ^[18]
- VisuMap kMeans Clustering ^[19]

Source code

- ELKI and Weka are written in Java and include k-means and variations
- K-means application in PHP, ^[20] using VB, ^[21] using Perl, ^[22] using C++, ^[23] using Matlab, ^[24] using Ruby, ^{[25][26]} using Python with scipy, ^[27] using X10^[28]
- A parallel out-of-core implementation in C^[29]
- An open-source collection of clustering algorithms, including k-means, implemented in Javascript.^[30] Online demo.^[31]

Visualization, animation and examples

- ELKI can visualize k-means using Voronoi cells and Delaunay triangulation for 2D data. In higher dimensionality, only cluster assignments and cluster centers are visualized
- Demos of the K-means-algorithm^{[32][33][34][35][36][37]}
- K-means and K-medoids (Applet), University of Leicester^[1]
- Clustergram - cluster diagnostic plot - for visual diagnostics of choosing the number of (k) clusters (R code)^[38]

References

- [1] http://en.wikipedia.org/w/index.php?title=Template:Machine_learning_bar&action=edit
- [2] Published in journal much later:
- [3] Since the square root is a monotone function, this also is the minimum Euclidean distance assignment.
- [4] Honarkhah, M and Caers, J, 2010, *Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling* (<http://dx.doi.org/10.1007/s11004-010-9276-7>), *Mathematical Geosciences*, 42: 487 - 517
- [5] <http://cwiki.apache.org/MAHOUT/k-means-clustering.html>
- [6] <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/kmeans.html>
- [7] <http://docs.scipy.org/doc/scipy/reference/cluster.vq.html>
- [8] <http://www.codeding.com/?article=14>
- [9] <http://pgxn.org/dist/kmeans/>
- [10] <http://graphlab.org/toolkits/clustering/>
- [11] <http://spectralpython.sourceforge.net/algorithms.html#k-means-clustering>
- [12] <http://scikit-learn.org/dev/modules/generated/sklearn.cluster.KMeans.html>
- [13] <http://docs.opencv.org/modules/core/doc/clustering.html?highlight=kmeans#cv2.kmeans>
- [14] <http://gforge.inria.fr/projects/yael/>
- [15] <http://reference.wolfram.com/mathematica/ref/ClusteringComponents.html>
- [16] <http://www.mathworks.com/access/helpdesk/help/toolbox/stats/kmeans.html>
- [17] http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/fastclus_toc.htm
- [18] <http://www.stata.com/help13.cgi?cluster+kmeans>
- [19] <http://www.visumap.com/index.aspx?p=Products>
- [20] <http://www25.brinkster.com/denshade/kmeans.php.htm>
- [21] K-Means Clustering Tutorial: Download (<http://people.revoledu.com/kardi/tutorial/kMean/download.htm>)
- [22] Perl script for Kmeans clustering (http://www.lwebzem.com/cgi-bin/k_means/test3.cgi)
- [23] Antonio Gulli's coding playground: K-means in C (<http://codingplayground.blogspot.com/2009/03/k-means-in-c.html>)
- [24] K-Means Clustering Tutorial: Matlab Code (http://people.revoledu.com/kardi/tutorial/kMean/matlab_kMeans.htm)
- [25] AI4R :: Artificial Intelligence for Ruby (<http://ai4r.org/index.html>)
- [26] reddavis/K-Means - GitHub (<http://github.com/reddavis/K-Means/tree/master>)

-
- [27] K-means clustering and vector quantization (scipy.cluster.vq) — SciPy v0.11 Reference Guide (DRAFT) (<http://docs.scipy.org/doc/scipy/reference/cluster.vq.html>)
 - [28] <http://dist.codehaus.org/x10/applications/samples/KMeansDist.x10>
 - [29] <http://www.cs.princeton.edu/~wdong/kmeans/>
 - [30] <http://code.google.com/p/figure/FIGUE>
 - [31] <http://jydelort.appspot.com/resources/figure/demo.html>
 - [32] Clustering - K-means demo (http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html)
 - [33] siebn.de - YAK-Means (<http://siebn.de/other/yakmeans/>)
 - [34] k-Means and Voronoi Tesselation: Built with Processing | Information & Visualization (<http://informationandvisualization.de/blog/kmeans-and-voronoi-tesselation-built-processing>)
 - [35] Hyper-threaded Java - JavaWorld (<http://www.javaworld.com/javaworld/jw-11-2006/jw-1121-thread.html>)
 - [36] Color clustering (<http://www.leet.it/home/lale/clustering/>)
 - [37] Interactive step-by-step examples in Javascript of good and bad k-means clustering (<http://www.onmyphd.com/?p=k-means.clustering>)
 - [38] Clustergram: visualization and diagnostics for cluster analysis (R code) | R-statistics blog (<http://www.r-statistics.com/2010/06/clustergram-visualization-and-diagnostics-for-cluster-analysis-r-code/>)
-

Matrix

Matrix (mathematics)

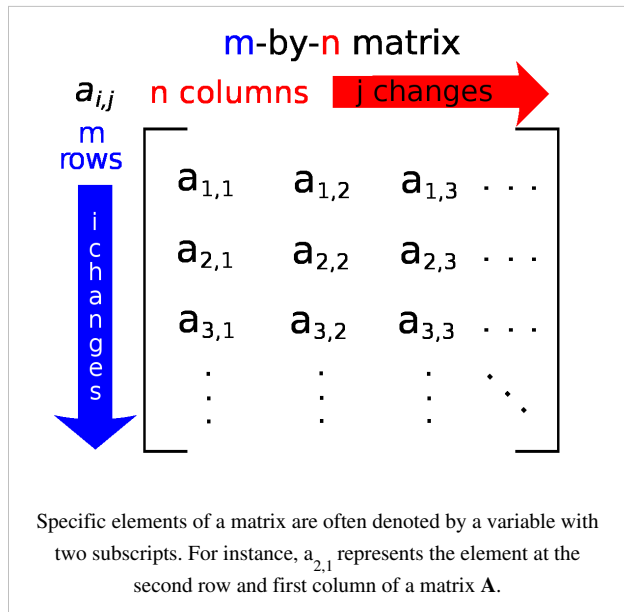
In mathematics, a **matrix** (plural **matrices**) is a rectangular *array*^[1] of numbers, symbols, or expressions, arranged in *rows* and *columns*. The individual items in a matrix are called its *elements* or *entries*. An example of a matrix with 2 rows and 3 columns is

$$\begin{bmatrix} 1 & 9 & -13 \\ 20 & 5 & -6 \end{bmatrix}.$$

Matrices of the same size can be added or subtracted element by element. But the rule for matrix multiplication is that two matrices can be multiplied only when the number of columns in the first equals the number of rows in the second. A major application of matrices is to represent linear transformations, that is, generalizations of linear functions such as $f(x) = 4x$. For example, the rotation of vectors in three dimensional space is a linear transformation. If \mathbf{R} is a rotation matrix and \mathbf{v} is a column vector (a matrix with only one column) describing the position of a point in space, the product $\mathbf{R}\mathbf{v}$ is a column vector describing the position of that point after a rotation. The product of two matrices is a matrix that represents the composition of two linear transformations. Another application of matrices is in the solution of a system of linear equations. If the matrix is square, it is possible to deduce some of its properties by computing its determinant. For example, a square matrix has an inverse if and only if its determinant is not zero. Eigenvalues and eigenvectors provide insight into the geometry of linear transformations.

Applications of matrices are found in most scientific fields. In every branch of physics, including classical mechanics, optics, electromagnetism, quantum mechanics, and quantum electrodynamics, they are used to study physical phenomena, such as the motion of rigid bodies. In computer graphics, they are used to project a 3-dimensional image onto a 2-dimensional screen. In probability theory and statistics, stochastic matrices are used to describe sets of probabilities; for instance, they are used within the PageRank algorithm that ranks the pages in a Google search.^[2] Matrix calculus generalizes classical analytical notions such as derivatives and exponentials to higher dimensions.

A major branch of numerical analysis is devoted to the development of efficient algorithms for matrix computations, a subject that is centuries old and is today an expanding area of research. Matrix decomposition methods simplify computations, both theoretically and practically. Algorithms that are tailored to particular matrix structures, such as sparse matrices and near-diagonal matrices, expedite computations in finite element method and other computations. Infinite matrices occur in planetary theory and in atomic theory. A simple example of an infinite matrix is the matrix representing the derivative operator, which acts on the Taylor series of a function.



Definition

A *matrix* is a rectangular array of numbers or other mathematical objects, for which operations such as addition and multiplication are defined. Most commonly, a matrix over a field F is a rectangular array of scalars from F . Most of this article focuses on *real* and *complex matrices*, i.e., matrices whose elements are real numbers or complex numbers, respectively. More general types of entries are discussed below. For instance, this is a real matrix:

$$\mathbf{A} = \begin{bmatrix} -1.3 & 0.6 \\ 20.4 & 5.5 \\ 9.7 & -6.2 \end{bmatrix}.$$

The numbers, symbols or expressions in the matrix are called its *entries* or its *elements*. The horizontal and vertical lines of entries in a matrix are called *rows* and *columns*, respectively.

Size

The size of a matrix is defined by the number of rows and columns that it contains. A matrix with m rows and n columns is called an $m \times n$ matrix or *m-by-n matrix*, while m and n are called its *dimensions*. For example, the matrix \mathbf{A} above is a 3×2 matrix.

Matrices which have a single row are called *row vectors*, and those which have a single column are called *column vectors*. A matrix which has the same number of rows and columns is called a *square matrix*. A matrix with an infinite number of rows or columns (or both) is called an *infinite matrix*. In some contexts such as computer algebra programs it is useful to consider a matrix with no rows or no columns, called an *empty matrix*.

Name	Size	Example	Description
Row vector	$1 \times n$	$[3 \ 7 \ 2]$	A matrix with one row, sometimes used to represent a vector
Column vector	$n \times 1$	$\begin{bmatrix} 4 \\ 1 \\ 8 \end{bmatrix}$	A matrix with one column, sometimes used to represent a vector
Square matrix	$n \times n$	$\begin{bmatrix} 9 & 13 & 5 \\ 1 & 11 & 7 \\ 2 & 6 & 3 \end{bmatrix}$	A matrix with the same number of rows and columns, sometimes used to represent a linear transformation from a vector space to itself, such as reflection, rotation, or shearing.

Notation

Matrices are commonly written in box brackets:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

An alternative notation uses large parentheses instead of box brackets:

$$\mathbf{A} = \left(\begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{array} \right).$$

The specifics of symbolic matrix notation varies widely, with some prevailing trends. Matrices are usually symbolized using upper-case letters (such as \mathbf{A} in the examples above), while the corresponding lower-case letters, with two subscript indices (e.g., a_{11} , or $a_{1,1}$), represent the entries. In addition to using upper-case letters to symbolize matrices, many authors use a special typographical style, commonly boldface upright (non-italic), to

further distinguish matrices from other mathematical objects. An alternative notation involves the use of a double-underline with the variable name, with or without boldface style, (e.g., $\underline{\underline{\mathbf{A}}}$).

The entry in the i -th row and j -th column of a matrix \mathbf{A} is sometimes referred to as the i_j , (i,j) , or $(i,j)^{\text{th}}$ entry of the matrix, and most commonly denoted as $a_{i,j}$, or a_{ij} . Alternative notations for that entry are $A[i,j]$ or $A_{i,j}$. For example, the (1,3) entry of the following matrix \mathbf{A} is 5 (also denoted a_{13} , $a_{1,3}$, $A[1,3]$ or $A_{1,3}$):

$$\mathbf{A} = \begin{bmatrix} 4 & -7 & 5 & 0 \\ -2 & 0 & 11 & 8 \\ 19 & 1 & -3 & 12 \end{bmatrix}$$

Sometimes, the entries of a matrix can be defined by a formula such as $a_{i,j} = f(i, j)$. For example, each of the entries of the following matrix \mathbf{A} is determined by $a_{ij} = i - j$.


$$\mathbf{A} = \begin{bmatrix} 0 & -1 & -2 & -3 \\ 1 & 0 & -1 & -2 \\ 2 & 1 & 0 & -1 \end{bmatrix}$$

In this case, the matrix itself is sometimes defined by that formula, within square brackets or double parenthesis. For example, the matrix above is defined as $\mathbf{A} = [i-j]$, or $\mathbf{A} = ((i-j))$. If matrix size is $m \times n$, the above-mentioned formula $f(i, j)$ is valid for any $i = 1, \dots, m$ and any $j = 1, \dots, n$. This can be either specified separately, or using $m \times n$ as a subscript. For instance, the matrix \mathbf{A} above is 3×4 and can be defined as $\mathbf{A} = [i - j]_{(i = 1, 2, 3; j = 1, \dots, 4)}$, or $\mathbf{A} = [i - j]_{3 \times 4}$.

Some programming languages utilize doubly-subscripted arrays (or arrays of arrays) to represent an $m \times n$ matrix. Some programming languages start the numbering of array indexes at zero, in which case the entries of an m -by- n matrix are indexed by $0 \leq i \leq m - 1$ and $0 \leq j \leq n - 1$. This article follows the more common convention in mathematical writing where enumeration starts from 1.

The set of all m -by- n matrices is denoted $\square(m, n)$.

Basic operations

 How to organize, add and multiply matrices - Bill Shillito ^[3], TED ED

There are a number of basic operations that can be applied to modify matrices, called *matrix addition*, *scalar multiplication*, *transposition*, *matrix multiplication*, *row operations*, and *submatrix*.

Addition, scalar multiplication and transposition

Operation	Definition	Example
Addition	The <i>sum</i> $\mathbf{A}+\mathbf{B}$ of two m -by- n matrices \mathbf{A} and \mathbf{B} is calculated entrywise: $(\mathbf{A} + \mathbf{B})_{i,j} = \mathbf{A}_{i,j} + \mathbf{B}_{i,j}$, where $1 \leq i \leq m$ and $1 \leq j \leq n$.	$\begin{bmatrix} 1 & 3 & 1 \\ 1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 5 \\ 7 & 5 & 0 \end{bmatrix} = \begin{bmatrix} 1+0 & 3+0 & 1+5 \\ 1+7 & 0+5 & 0+0 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 6 \\ 8 & 5 & 0 \end{bmatrix}$
Scalar multiplication	The <i>scalar multiplication</i> $c\mathbf{A}$ of a matrix \mathbf{A} and a number c (also called a scalar in the parlance of abstract algebra) is given by multiplying every entry of \mathbf{A} by c : $(c\mathbf{A})_{i,j} = c \cdot \mathbf{A}_{i,j}$.	$2 \cdot \begin{bmatrix} 1 & 8 & -3 \\ 4 & -2 & 5 \end{bmatrix} = \begin{bmatrix} 2 \cdot 1 & 2 \cdot 8 & 2 \cdot -3 \\ 2 \cdot 4 & 2 \cdot -2 & 2 \cdot 5 \end{bmatrix} = \begin{bmatrix} 2 & 16 & -6 \\ 8 & -4 & 10 \end{bmatrix}$
Transpose	The <i>transpose</i> of an m -by- n matrix \mathbf{A} is the n -by- m matrix \mathbf{A}^T (also denoted \mathbf{A}^{tr} or ${}^t\mathbf{A}$) formed by turning rows into columns and vice versa: $(\mathbf{A}^T)_{i,j} = \mathbf{A}_{j,i}$.	$\begin{bmatrix} 1 & 2 & 3 \\ 0 & -6 & 7 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 \\ 2 & -6 \\ 3 & 7 \end{bmatrix}$

Familiar properties of numbers extend to these operations of matrices: for example, addition is commutative, i.e., the matrix sum does not depend on the order of the summands: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$. The transpose is compatible with addition and scalar multiplication, as expressed by $(c\mathbf{A})^T = c(\mathbf{A}^T)$ and $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$. Finally, $(\mathbf{A}^T)^T = \mathbf{A}$.

Matrix multiplication

Multiplication of two matrices is defined only if the number of columns of the left matrix is the same as the number of rows of the right matrix. If \mathbf{A} is an m -by- n matrix and \mathbf{B} is an n -by- p matrix, then their *matrix product* \mathbf{AB} is the m -by- p matrix whose entries are given by dot product of the corresponding row of \mathbf{A} and the corresponding column of \mathbf{B} :

$$[\mathbf{AB}]_{i,j} = A_{i,1}B_{1,j} + A_{i,2}B_{2,j} + \dots + A_{i,n}B_{n,j} = \sum_{r=1}^n A_{i,r}B_{r,j},$$

where $1 \leq i \leq m$ and $1 \leq j \leq p$. For example, the underlined entry 2340 in the product is calculated as $(2 \times 1000) + (3 \times 100) + (4 \times 10) = 2340$:

$$\begin{bmatrix} \underline{2} & \underline{3} & \underline{4} \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & \underline{1000} \\ 1 & \underline{100} \\ 0 & \underline{10} \end{bmatrix} = \begin{bmatrix} \underline{3} & \underline{2340} \\ 0 & 1000 \end{bmatrix}.$$

Matrix multiplication satisfies the rules $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ (associativity), and $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$ as well as $\mathbf{C}(\mathbf{A}+\mathbf{B}) = \mathbf{CA}+\mathbf{CB}$ (left and right distributivity), whenever the size of the matrices is such that the various products are defined. The product \mathbf{AB} may be defined without \mathbf{BA} being defined, namely if \mathbf{A} and \mathbf{B} are m -by- n and n -by- k matrices, respectively, and $m \neq k$. Even if both products are defined, they need not be equal, i.e., generally one has

$$\mathbf{AB} \neq \mathbf{BA},$$

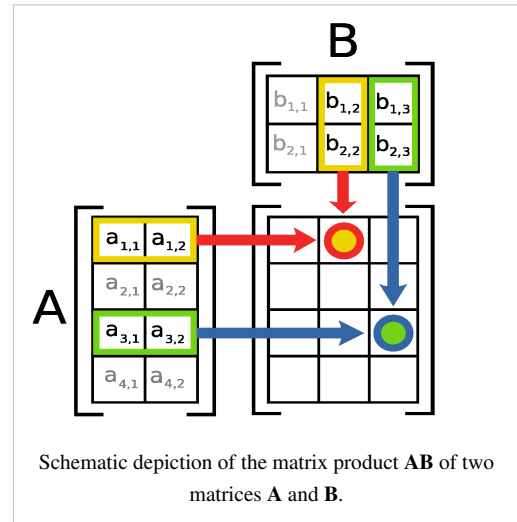
i.e., matrix multiplication is not commutative, in marked contrast to (rational, real, or complex) numbers whose product is independent of the order of the factors. An example of two matrices not commuting with each other is:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 3 \end{bmatrix},$$

whereas

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 0 & 0 \end{bmatrix}.$$

Besides the ordinary matrix multiplication just described, there exist other less frequently used operations on matrices that can be considered forms of multiplication, such as the Hadamard product and the Kronecker product. They arise in solving matrix equations such as the Sylvester equation.



Row operations

There are three types of row operations:

1. row addition, that is adding a row to another.
2. row multiplication, that is multiplying all entries of a row by a non-zero constant;
3. row switching, that is interchanging two rows of a matrix;

These operations are used in a number of ways, including solving linear equations and finding matrix inverses.

Submatrix

A **submatrix** of a matrix is obtained by deleting any collection of rows and/or columns. For example, for the following 3-by-4 matrix, we can construct a 2-by-3 submatrix by removing row 3 and column 2:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 3 & 4 \\ 5 & 7 & 8 \end{bmatrix}.$$

The minors and cofactors of a matrix are found by computing the determinant of certain submatrices.

Linear equations

Matrices can be used to compactly write and work with multiple linear equations, i.e., systems of linear equations. For example, if \mathbf{A} is an m -by- n matrix, \mathbf{x} designates a column vector (i.e., $n \times 1$ -matrix) of n variables x_1, x_2, \dots, x_n , and \mathbf{b} is an $m \times 1$ -column vector, then the matrix equation

$$\mathbf{Ax} = \mathbf{b}$$

is equivalent to the system of linear equations

$$A_{1,1}x_1 + A_{1,2}x_2 + \dots + A_{1,n}x_n = b_1$$

...

$$A_{m,1}x_1 + A_{m,2}x_2 + \dots + A_{m,n}x_n = b_m.$$

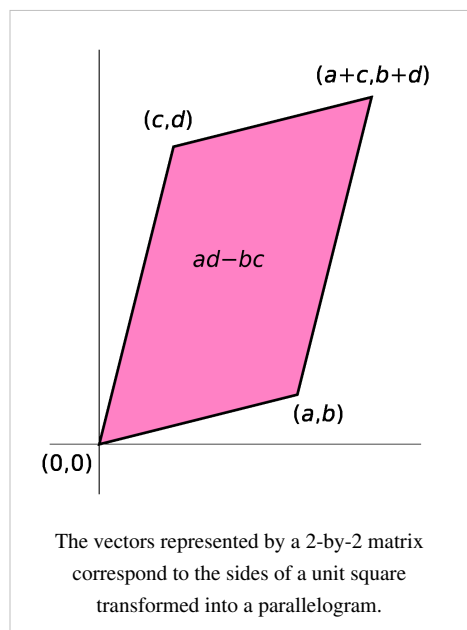
Linear transformations

Matrices and matrix multiplication reveal their essential features when related to *linear transformations*, also known as *linear maps*. A real m -by- n matrix \mathbf{A} gives rise to a linear transformation $\mathbf{R}^n \rightarrow \mathbf{R}^m$ mapping each vector \mathbf{x} in \mathbf{R}^n to the (matrix) product \mathbf{Ax} , which is a vector in \mathbf{R}^m . Conversely, each linear transformation $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ arises from a unique m -by- n matrix \mathbf{A} : explicitly, the (i, j) -entry of \mathbf{A} is the i^{th} coordinate of $f(\mathbf{e}_j)$, where $\mathbf{e}_j = (0, \dots, 0, 1, 0, \dots, 0)$ is the unit vector with 1 in the j^{th} position and 0 elsewhere. The matrix \mathbf{A} is said to represent the linear map f , and \mathbf{A} is called the *transformation matrix* of f .

For example, the 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

can be viewed as the transform of the unit square into a parallelogram with vertices at $(0, 0)$, (a, b) , $(a + c, b + d)$, and (c, d) . The parallelogram pictured at the right is obtained by multiplying \mathbf{A} with



each of the column vectors $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ in turn. These vectors define the vertices of the unit square.

The following table shows a number of 2-by-2 matrices with the associated linear maps of \mathbf{R}^2 . The blue original is mapped to the green grid and shapes. The origin (0,0) is marked with a black point.

Horizontal shear with m=1.25.	Horizontal flip	Squeeze mapping with r=3/2	Scaling by a factor of 3/2	Rotation by $\pi/6^R = 30^\circ$
$\begin{bmatrix} 1 & 1.25 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 3/2 & 0 \\ 0 & 2/3 \end{bmatrix}$	$\begin{bmatrix} 3/2 & 0 \\ 0 & 3/2 \end{bmatrix}$	$\begin{bmatrix} \cos(\pi/6^R) & -\sin(\pi/6^R) \\ \sin(\pi/6^R) & \cos(\pi/6^R) \end{bmatrix}$

Under the 1-to-1 correspondence between matrices and linear maps, matrix multiplication corresponds to composition of maps: if a k -by- m matrix \mathbf{B} represents another linear map $g : \mathbf{R}^m \rightarrow \mathbf{R}^k$, then the composition $g \circ f$ is represented by \mathbf{BA} since

$$(g \circ f)(\mathbf{x}) = g(f(\mathbf{x})) = g(\mathbf{Ax}) = \mathbf{B}(\mathbf{Ax}) = (\mathbf{BA})\mathbf{x}.$$

The last equality follows from the above-mentioned associativity of matrix multiplication.

The rank of a matrix \mathbf{A} is the maximum number of linearly independent row vectors of the matrix, which is the same as the maximum number of linearly independent column vectors. Equivalently it is the dimension of the image of the linear map represented by \mathbf{A} . The rank-nullity theorem states that the dimension of the kernel of a matrix plus the rank equals the number of columns of the matrix.

Square matrices

A square matrix is a matrix with the same number of rows and columns. An n -by- n matrix is known as a square matrix of order n . Any two square matrices of the same order can be added and multiplied. The entries a_{ii} form the main diagonal of a square matrix. They lie on the imaginary line which runs from the top left corner to the bottom right corner of the matrix.

Main types

Name	Example with $n = 3$
Diagonal matrix	$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$
Lower triangular matrix	$\begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$
Upper triangular matrix	$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$

Diagonal and triangular matrices

If all entries of \mathbf{A} below the main diagonal are zero, \mathbf{A} is called an *upper triangular matrix*. Similarly if all entries of \mathbf{A} above the main diagonal are zero, \mathbf{A} is called a *lower triangular matrix*. If all entries outside the main diagonal are zero, \mathbf{A} is called a diagonal matrix.

Identity matrix

The identity matrix \mathbf{I}_n of size n is the n -by- n matrix in which all the elements on the main diagonal are equal to 1 and all other elements are equal to 0, e.g.

$$I_1 = [1], I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \dots, I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

It is a square matrix of order n , and also a special kind of diagonal matrix. It is called identity matrix because multiplication with it leaves a matrix unchanged:

$$\mathbf{A}\mathbf{I}_n = \mathbf{I}_m\mathbf{A} = \mathbf{A} \text{ for any } m\text{-by-}n \text{ matrix } \mathbf{A}.$$

Symmetric or skew-symmetric matrix

A square matrix \mathbf{A} that is equal to its transpose, i.e., $\mathbf{A} = \mathbf{A}^T$, is a symmetric matrix. If instead, \mathbf{A} was equal to the negative of its transpose, i.e., $\mathbf{A} = -\mathbf{A}^T$, then \mathbf{A} is a skew-symmetric matrix. In complex matrices, symmetry is often replaced by the concept of Hermitian matrices, which satisfy $\mathbf{A}^* = \mathbf{A}$, where the star or asterisk denotes the conjugate transpose of the matrix, i.e., the transpose of the complex conjugate of \mathbf{A} .

By the spectral theorem, real symmetric matrices and complex Hermitian matrices have an eigenbasis; i.e., every vector is expressible as a linear combination of eigenvectors. In both cases, all eigenvalues are real. This theorem can be generalized to infinite-dimensional situations related to matrices with infinitely many rows and columns, see below.

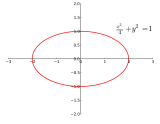
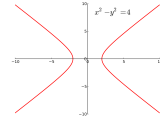
Invertible matrix and its inverse

A square matrix \mathbf{A} is called *invertible* or *non-singular* if there exists a matrix \mathbf{B} such that

$$\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A} = \mathbf{I}_n.$$

If \mathbf{B} exists, it is unique and is called the *inverse matrix* of \mathbf{A} , denoted \mathbf{A}^{-1} .

Definite matrix

Positive definite matrix	Indefinite matrix
$\begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1/4 & 0 \\ 0 & -1/4 \end{bmatrix}$
$Q(x,y) = 1/4 x^2 + y^2$	$Q(x,y) = 1/4 x^2 - 1/4 y^2$
	
Points such that $Q(x,y)=1$ (Ellipse).	Points such that $Q(x,y)=1$ (Hyperbola).

A symmetric $n \times n$ -matrix is called *positive-definite* (respectively *negative-definite*; *indefinite*), if for all nonzero vectors $\mathbf{x} \in \mathbf{R}^n$ the associated quadratic form given by

$$Q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$$

takes only positive values (respectively only negative values; both some negative and some positive values). If the quadratic form takes only non-negative (respectively only non-positive) values, the symmetric matrix is called positive-semidefinite (respectively negative-semidefinite); hence the matrix is indefinite precisely when it is neither positive-semidefinite nor negative-semidefinite.

A symmetric matrix is positive-definite if and only if all its eigenvalues are positive. The table at the right shows two possibilities for 2-by-2 matrices.

Allowing as input two different vectors instead yields the bilinear form associated to \mathbf{A} :

$$B_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{A} \mathbf{y}.$$

Orthogonal matrix

An *orthogonal matrix* is a square matrix with real entries whose columns and rows are orthogonal unit vectors (i.e., orthonormal vectors). Equivalently, a matrix A is orthogonal if its transpose is equal to its inverse:

$$A^T = A^{-1},$$

which entails

$$A^T A = A A^T = I,$$

where I is the identity matrix.

An orthogonal matrix A is necessarily invertible (with inverse $A^{-1} = A^T$), unitary ($A^{-1} = A^*$), and normal ($A^*A = AA^*$). The determinant of any orthogonal matrix is either +1 or -1. A *special orthogonal matrix* is an orthogonal matrix with determinant +1. As a linear transformation, every orthogonal matrix with determinant +1 is a pure rotation, while every orthogonal matrix with determinant -1 is either a pure reflection, or a composition of reflection and rotation.

The complex analogue of an orthogonal matrix is a unitary matrix.

Main operations

Trace

The trace, $\text{tr}(\mathbf{A})$ of a square matrix \mathbf{A} is the sum of its diagonal entries. While matrix multiplication is not commutative as mentioned above, the trace of the product of two matrices is independent of the order of the factors:

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}).$$

This is immediate from the definition of matrix multiplication:

$$\text{tr}(\mathbf{AB}) = \sum_{i=1}^m \sum_{j=1}^n A_{ij} B_{ji} = \text{tr}(\mathbf{BA}).$$

Also, the trace of a matrix is equal to that of its transpose, i.e.,

$$\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^T).$$

Determinant

The *determinant* $\det(\mathbf{A})$ or $|\mathbf{A}|$ of a square matrix \mathbf{A} is a number encoding certain properties of the matrix. A matrix is invertible if and only if its determinant is nonzero. Its absolute value equals the area (in \mathbf{R}^2) or volume (in \mathbf{R}^3) of the image of the unit square (or cube), while its sign corresponds to the orientation of the corresponding linear map: the determinant is positive if and only if the orientation is preserved.

The determinant of 2-by-2 matrices is given by

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

The determinant of 3-by-3 matrices involves 6 terms (rule of Sarrus). The more lengthy Leibniz formula generalises these two formulae to all dimensions.

The determinant of a product of square matrices equals the product of their determinants:

$$\det(\mathbf{AB}) = \det(\mathbf{A}) \cdot \det(\mathbf{B}).$$

Adding a multiple of any row to another row, or a multiple of any column to another column, does not change the determinant. Interchanging two rows or two columns affects the determinant by multiplying it by -1 . Using these operations, any matrix can be transformed to a lower (or upper) triangular matrix, and for such matrices the determinant equals the product of the entries on the main diagonal; this provides a method to calculate the determinant of any matrix. Finally, the Laplace expansion expresses the determinant in terms of minors, i.e., determinants of smaller matrices. This expansion can be used for a recursive definition of determinants (taking as starting case the determinant of a 1-by-1 matrix, which is its unique entry, or even the determinant of a 0-by-0 matrix, which is 1), that can be seen to be equivalent to the Leibniz formula. Determinants can be used to solve linear systems using Cramer's rule, where the division of the determinants of two related square matrices equates to the value of each of the system's variables.

Eigenvalues and eigenvectors

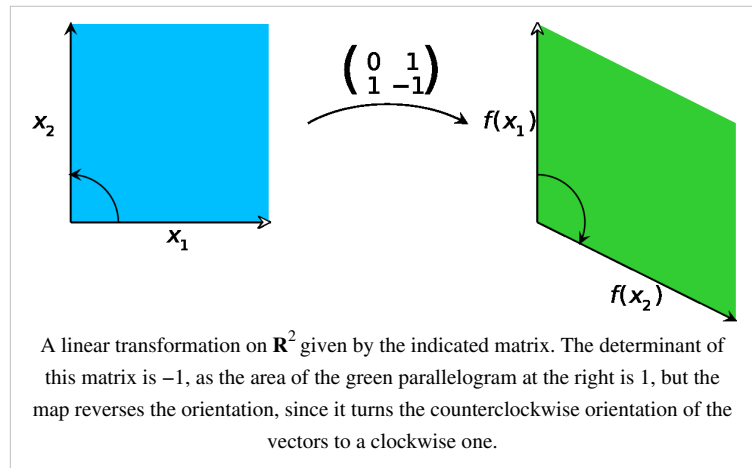
A number λ and a non-zero vector \mathbf{v} satisfying

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

are called an *eigenvalue* and an *eigenvector* of \mathbf{A} , respectively.^[4] The number λ is an eigenvalue of an $n \times n$ -matrix \mathbf{A} if and only if $\mathbf{A} - \lambda\mathbf{I}_n$ is not invertible, which is equivalent to

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0.$$

The polynomial $p_{\mathbf{A}}$ in an indeterminate X given by evaluation the determinant $\det(X\mathbf{I}_n - \mathbf{A})$ is called the characteristic polynomial of \mathbf{A} . It is a monic polynomial of degree n . Therefore the polynomial equation $p_{\mathbf{A}}(\lambda) = 0$ has at most n different solutions, i.e., eigenvalues of the matrix. They may be complex even if the entries of \mathbf{A} are real. According to the Cayley–Hamilton theorem, $p_{\mathbf{A}}(\mathbf{A}) = \mathbf{0}$, that is, the result of substituting the matrix itself into its own characteristic polynomial yields the zero matrix.



Computational aspects

Matrix calculations can be often performed with different techniques. Many problems can be solved by both direct algorithms or iterative approaches. For example, the eigenvectors of a square matrix can be obtained by finding a sequence of vectors \mathbf{x}_n converging to an eigenvector when n tends to infinity.

To be able to choose the more appropriate algorithm for each specific problem, it is important to determine both the effectiveness and precision of all the available algorithms. The domain studying these matters is called numerical linear algebra. As with other numerical situations, two main aspects are the complexity of algorithms and their numerical stability.

Determining the complexity of an algorithm means finding upper bounds or estimates of how many elementary operations such as additions and multiplications of scalars are necessary to perform some algorithm, e.g., multiplication of matrices. For example, calculating the matrix product of two n -by- n matrix using the definition given above needs n^3 multiplications, since for any of the n^2 entries of the product, n multiplications are necessary. The Strassen algorithm outperforms this "naive" algorithm; it needs only $n^{2.807}$ multiplications. A refined approach also incorporates specific features of the computing devices.

In many practical situations additional information about the matrices involved is known. An important case are sparse matrices, i.e., matrices most of whose entries are zero. There are specifically adapted algorithms for, say, solving linear systems $\mathbf{Ax} = \mathbf{b}$ for sparse matrices \mathbf{A} , such as the conjugate gradient method.

An algorithm is, roughly speaking, numerically stable, if little deviations in the input values do not lead to big deviations in the result. For example, calculating the inverse of a matrix via Laplace's formula ($\text{Adj}(\mathbf{A})$ denotes the adjugate matrix of \mathbf{A})

$$\mathbf{A}^{-1} = \text{Adj}(\mathbf{A}) / \det(\mathbf{A})$$

may lead to significant rounding errors if the determinant of the matrix is very small. The norm of a matrix can be used to capture the conditioning of linear algebraic problems, such as computing a matrix' inverse.

Although most computer languages are not designed with commands or libraries for matrices, as early as the 1970s, some engineering desktop computers such as the HP 9830 had ROM cartridges to add BASIC commands for matrices. Some computer languages such as APL were designed to manipulate matrices, and various mathematical programs can be used to aid computing with matrices.^[5]

Decomposition

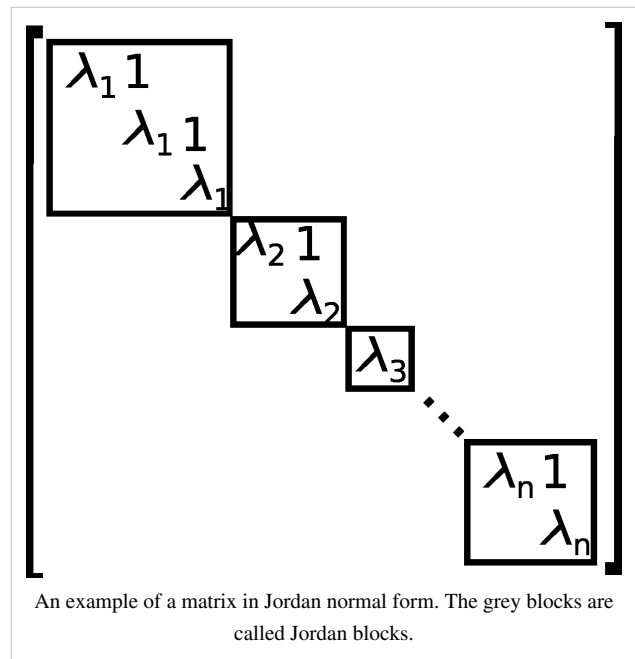
There are several methods to render matrices into a more easily accessible form. They are generally referred to as *matrix decomposition* or *matrix factorization* techniques. The interest of all these techniques is that they preserve certain properties of the matrices in question, such as determinant, rank or inverse, so that these quantities can be calculated after applying the transformation, or that certain matrix operations are algorithmically easier to carry out for some types of matrices.

The LU decomposition factors matrices as a product of lower (\mathbf{L}) and an upper triangular matrices (\mathbf{U}). Once this decomposition is calculated, linear systems can be solved more efficiently, by a simple technique called forward and back substitution. Likewise, inverses of triangular matrices are algorithmically easier to calculate. The *Gaussian elimination* is a similar algorithm; it transforms any matrix to row echelon form. Both methods proceed by multiplying the matrix by suitable elementary matrices, which correspond to permuting rows or columns and adding multiples of one row to another row. Singular value decomposition expresses any matrix \mathbf{A} as a product \mathbf{UDV}^* , where \mathbf{U} and \mathbf{V} are unitary matrices and \mathbf{D} is a diagonal matrix.

The eigendecomposition or *diagonalization* expresses \mathbf{A} as a product \mathbf{VDV}^{-1} , where \mathbf{D} is a diagonal matrix and \mathbf{V} is a suitable invertible matrix. If \mathbf{A} can be written in this form, it is called diagonalizable. More generally, and applicable to all matrices, the Jordan decomposition transforms a matrix into Jordan normal form, that is to say matrices whose only nonzero entries are the eigenvalues λ_1 to λ_n of \mathbf{A} , placed on the main diagonal and possibly entries equal to one directly above the main diagonal, as shown at the right. Given the eigendecomposition, the n^{th} power of \mathbf{A} (i.e., n -fold iterated matrix multiplication) can be calculated via

$$\mathbf{A}^n = (\mathbf{VDV}^{-1})^n = \mathbf{VDV}^{-1}\mathbf{VDV}^{-1}\dots\mathbf{VDV}^{-1} = \mathbf{VD}^n\mathbf{V}^{-1}$$

and the power of a diagonal matrix can be calculated by taking the corresponding powers of the diagonal entries, which is much easier than doing the exponentiation for \mathbf{A} instead. This can be used to compute the matrix exponential $e^{\mathbf{A}}$, a need frequently arising in solving linear differential equations, matrix logarithms and square roots of matrices. To avoid numerically ill-conditioned situations, further algorithms such as the Schur decomposition can be employed.



Abstract algebraic aspects and generalizations

Matrices can be generalized in different ways. Abstract algebra uses matrices with entries in more general fields or even rings, while linear algebra codifies properties of matrices in the notion of linear maps. It is possible to consider matrices with infinitely many columns and rows. Another extension are tensors, which can be seen as higher-dimensional arrays of numbers, as opposed to vectors, which can often be realised as sequences of numbers, while matrices are rectangular or two-dimensional array of numbers. Matrices, subject to certain requirements tend to form groups known as matrix groups.

Matrices with more general entries

This article focuses on matrices whose entries are real or complex numbers. However, matrices can be considered with much more general types of entries than real or complex numbers. As a first step of generalization, any field, i.e., a set where addition, subtraction, multiplication and division operations are defined and well-behaved, may be used instead of \mathbf{R} or \mathbf{C} , for example rational numbers or finite fields. For example, coding theory makes use of matrices over finite fields. Wherever eigenvalues are considered, as these are roots of a polynomial they may exist only in a larger field than that of the coefficients of the matrix; for instance they may be complex in case of a matrix with real entries. The possibility to reinterpret the entries of a matrix as elements of a larger field (e.g., to view a real matrix as a complex matrix whose entries happen to be all real) then allows considering each square matrix to possess a full set of eigenvalues. Alternatively one can consider only matrices with entries in an algebraically closed field, such as \mathbf{C} , from the outset.

More generally, abstract algebra makes great use of matrices with entries in a ring R . Rings are a more general notion than fields in that a division operation need not exist. The very same addition and multiplication operations of matrices extend to this setting, too. The set $M(n, R)$ of all square n -by- n matrices over R is a ring called matrix ring, isomorphic to the endomorphism ring of the left R -module R^n . If the ring R is commutative, i.e., its multiplication is commutative, then $M(n, R)$ is a unitary noncommutative (unless $n = 1$) associative algebra over R . The determinant

of square matrices over a commutative ring R can still be defined using the Leibniz formula; such a matrix is invertible if and only if its determinant is invertible in R , generalising the situation over a field F , where every nonzero element is invertible. Matrices over superrings are called supermatrices.

Matrices do not always have all their entries in the same ring – or even in any ring at all. One special but common case is block matrices, which may be considered as matrices whose entries themselves are matrices. The entries need not be quadratic matrices, and thus need not be members of any ordinary ring; but their sizes must fulfil certain compatibility conditions.

Relationship to linear maps

Linear maps $\mathbf{R}^n \rightarrow \mathbf{R}^m$ are equivalent to m -by- n matrices, as described above. More generally, any linear map $f: V \rightarrow W$ between finite-dimensional vector spaces can be described by a matrix $\mathbf{A} = (a_{ij})$, after choosing bases $\mathbf{v}_1, \dots, \mathbf{v}_n$ of V , and $\mathbf{w}_1, \dots, \mathbf{w}_m$ of W (so n is the dimension of V and m is the dimension of W), which is such that

$$f(\mathbf{v}_j) = \sum_{i=1}^m a_{i,j} \mathbf{w}_i \quad \text{for } j = 1, \dots, n.$$

In other words, column j of A expresses the image of \mathbf{v}_j in terms of the basis vectors \mathbf{w}_i of W ; thus this relation uniquely determines the entries of the matrix \mathbf{A} . Note that the matrix depends on the choice of the bases: different choices of bases give rise to different, but equivalent matrices. Many of the above concrete notions can be reinterpreted in this light, for example, the transpose matrix \mathbf{A}^T describes the transpose of the linear map given by \mathbf{A} , with respect to the dual bases.

These properties can be restated in a more natural way: the category of all matrices with entries in a field k with multiplication as composition is equivalent to the category of finite dimensional vector spaces and linear maps over this field.

More generally, the set of $m \times n$ matrices can be used to represent the R -linear maps between the free modules R^m and R^n for an arbitrary ring R with unity. When $n = m$ composition of these maps is possible, and this gives rise to the matrix ring of $n \times n$ matrices representing the endomorphism ring of R^n .

Matrix groups

A group is a mathematical structure consisting of a set of objects together with a binary operation, i.e., an operation combining any two objects to a third, subject to certain requirements.^[6] A group in which the objects are matrices and the group operation is matrix multiplication is called a *matrix group*.^[7] Since in a group every element has to be invertible, the most general matrix groups are the groups of all invertible matrices of a given size, called the general linear groups.

Any property of matrices that is preserved under matrix products and inverses can be used to define further matrix groups. For example, matrices with a given size and with a determinant of 1 form a subgroup of (i.e., a smaller group contained in) their general linear group, called a special linear group. Orthogonal matrices, determined by the condition

$$\mathbf{M}^T \mathbf{M} = \mathbf{I},$$

form the orthogonal group. Every orthogonal matrix has determinant 1 or -1 . Orthogonal matrices with determinant 1 form a subgroup called *special orthogonal group*.

Every finite group is isomorphic to a matrix group, as one can see by considering the regular representation of the symmetric group. General groups can be studied using matrix groups, which are comparatively well-understood, by means of representation theory.^[8]

Infinite matrices

It is also possible to consider matrices with infinitely many rows and/or columns^[9] even if, being infinite objects, one cannot write down such matrices explicitly. All that matters is that for every element in the set indexing rows, and every element in the set indexing columns, there is a well-defined entry (these index sets need not even be subsets of the natural numbers). The basic operations of addition, subtraction, scalar multiplication and transposition can still be defined without problem; however matrix multiplication may involve infinite summations to define the resulting entries, and these are not defined in general.

If R is any ring with unity, then the ring of endomorphisms of $M = \bigoplus_{i \in I} R$ as a right R module is isomorphic to the ring of **column finite matrices** $\text{CFM}_I(R)$ whose entries are indexed by $I \times I$, and whose columns each contain only finitely many nonzero entries. The endomorphisms of M considered as a left R module result in an analogous object, the **row finite matrices** $\text{RFM}_I(R)$ whose rows each only have finitely many nonzero entries. If infinite matrices are used to describe linear maps, then only those matrices can be used all of whose columns have but a finite number of nonzero entries, for the following reason. For a matrix \mathbf{A} to describe a linear map $f: V \rightarrow W$, bases for both spaces must have been chosen; recall that by definition this means that every vector in the space can be written uniquely as a (finite) linear combination of basis vectors, so that written as a (column) vector v of coefficients, only finitely many entries v_i are nonzero. Now the columns of \mathbf{A} describe the images by f of individual basis vectors of V in the basis of W , which is only meaningful if these columns have only finitely many nonzero entries. There is no restriction on the rows of \mathbf{A} however: in the product $\mathbf{A} \cdot v$ there are only finitely many nonzero coefficients of v involved, so every one of its entries, even if it is given as an infinite sum of products, involves only finitely many nonzero terms and is therefore well defined. Moreover this amounts to forming a linear combination of the columns of \mathbf{A} that effectively involves only finitely many of them, whence the result has only finitely many nonzero entries, because each of those columns do. One also sees that products of two matrices of the given type is well defined (provided as usual that the column-index and row-index sets match), is again of the same type, and corresponds to the composition of linear maps.

If R is a normed ring, then the condition of row or column finiteness can be relaxed. With the norm in place, absolutely convergent series can be used instead of finite sums. For example, the matrices whose column sums are absolutely convergent sequences form a ring. Analogously of course, the matrices whose row sums are absolutely convergent series also form a ring.

In that vein, infinite matrices can also be used to describe operators on Hilbert spaces, where convergence and continuity questions arise, which again results in certain constraints that have to be imposed. However, the explicit point of view of matrices tends to obfuscate the matter,^[10] and the abstract and more powerful tools of functional analysis can be used instead.

Empty matrices

An *empty matrix* is a matrix in which the number of rows or columns (or both) is zero.^{[11][12]} Empty matrices help dealing with maps involving the zero vector space. For example, if A is a 3-by-0 matrix and B is a 0-by-3 matrix, then AB is the 3-by-3 zero matrix corresponding to the null map from a 3-dimensional space V to itself, while BA is a 0-by-0 matrix. There is no common notation for empty matrices, but most computer algebra systems allow creating and computing with them. The determinant of the 0-by-0 matrix is 1 as follows from regarding the empty product occurring in the Leibniz formula for the determinant as 1. This value is also consistent with the fact that the identity map from any finite dimensional space to itself has determinant 1, a fact that is often used as a part of the characterization of determinants.

Applications

There are numerous applications of matrices, both in mathematics and other sciences. Some of them merely take advantage of the compact representation of a set of numbers in a matrix. For example, in game theory and economics, the payoff matrix encodes the payoff for two players, depending on which out of a given (finite) set of alternatives the players choose. Text mining and automated thesaurus compilation makes use of document-term matrices such as tf-idf to track frequencies of certain words in several documents.

Complex numbers can be represented by particular real 2-by-2 matrices via

$$a + ib \leftrightarrow \begin{bmatrix} a & -b \\ b & a \end{bmatrix},$$

under which addition and multiplication of complex numbers and matrices correspond to each other. For example, 2-by-2 rotation matrices represent the multiplication with some complex number of absolute value 1, as above. A similar interpretation is possible for quaternions, and also for Clifford algebras in general.

Early encryption techniques such as the Hill cipher also used matrices. However, due to the linear nature of matrices, these codes are comparatively easy to break. Computer graphics uses matrices both to represent objects and to calculate transformations of objects using affine rotation matrices to accomplish tasks such as projecting a three-dimensional object onto a two-dimensional screen, corresponding to a theoretical camera observation. Matrices over a polynomial ring are important in the study of control theory.

Chemistry makes use of matrices in various ways, particularly since the use of quantum theory to discuss molecular bonding and spectroscopy. Examples are the overlap matrix and the Fock matrix used in solving the Roothaan equations to obtain the molecular orbitals of the Hartree–Fock method.

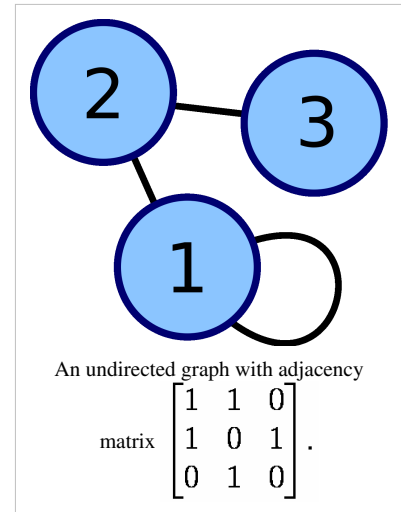
Graph theory

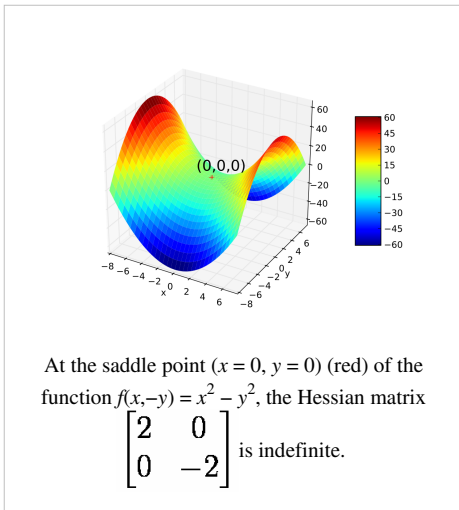
The adjacency matrix of a finite graph is a basic notion of graph theory. It saves which vertices of the graph are connected by an edge. Matrices containing just two different values (1 and 0 meaning for example "yes" and "no", respectively) are called logical matrices. The distance (or cost) matrix contains information about distances of the edges. These concepts can be applied to websites connected hyperlinks or cities connected by roads etc., in which case (unless the road network is extremely dense) the matrices tend to be sparse, i.e., contain few nonzero entries. Therefore, specifically tailored matrix algorithms can be used in network theory.

Analysis and geometry

The Hessian matrix of a differentiable function $f: \mathbf{R}^n \rightarrow \mathbf{R}$ consists of the second derivatives of f with respect to the several coordinate directions, i.e.

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_i \partial x_j} \end{bmatrix}.$$





It encodes information about the local growth behaviour of the function: given a critical point $\mathbf{x} = (x_1, \dots, x_n)$, i.e., a point where the first partial derivatives $\partial f / \partial x_i$ of f vanish, the function has a local minimum if the Hessian matrix is positive definite. Quadratic programming can be used to find global minima or maxima of quadratic functions closely related to the ones attached to matrices (see above).

Another matrix frequently used in geometrical situations is the Jacobi matrix of a differentiable map $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$. If f_1, \dots, f_m denote the components of f , then the Jacobi matrix is defined as

$$J(f) = \left[\frac{\partial f_i}{\partial x_j} \right]_{1 \leq i \leq m, 1 \leq j \leq n}$$

If $n > m$, and if the rank of the Jacobi matrix attains its maximal value m , f is locally invertible at that point, by the implicit function theorem.^[13]

Partial differential equations can be classified by considering the matrix of coefficients of the highest-order differential operators of the equation. For elliptic partial differential equations this matrix is positive definite, which has decisive influence on the set of possible solutions of the equation in question.

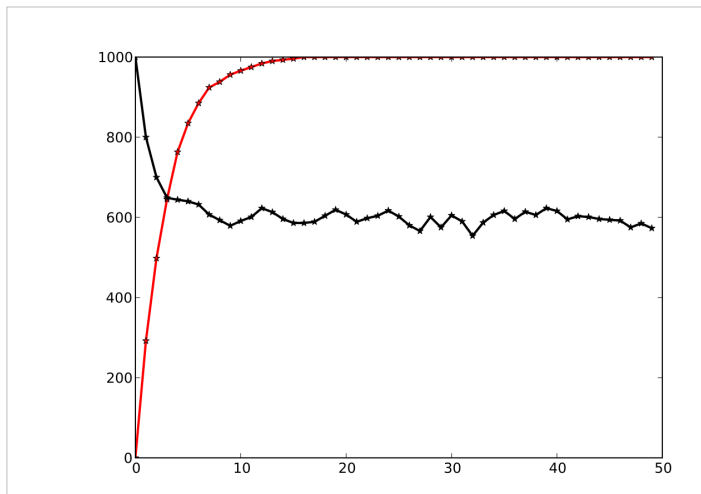
The finite element method is an important numerical method to solve partial differential equations, widely applied in simulating complex physical systems. It attempts to approximate the solution to some equation by piecewise linear functions, where the pieces are chosen with respect to a sufficiently fine grid, which in turn can be recast as a matrix equation.^[14]

Probability theory and statistics

Stochastic matrices are square matrices whose rows are probability vectors, i.e., whose entries are non-negative and sum up to one. Stochastic matrices are used to define Markov chains with finitely many states. A row of the stochastic matrix gives the probability distribution for the next position of some particle currently in the state that corresponds to the row. Properties of the Markov chain like absorbing states, i.e., states that any particle attains eventually, can be read off the eigenvectors of the transition matrices.

Statistics also makes use of matrices in many different forms. Descriptive statistics is concerned with describing data sets, which can often be represented as data matrices, which may then be subjected to dimensionality reduction techniques. The covariance matrix encodes the mutual variance of several random variables. Another technique using matrices are linear least squares, a method that approximates a finite set of pairs $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, by a linear function

$$y_i \approx ax_i + b, i = 1, \dots, N$$



Two different Markov chains. The chart depicts the number of particles (of a total of 1000) in state "2". Both limiting values can be determined from the transition matrices, which are given by $\begin{bmatrix} .7 & 0 \\ .3 & 1 \end{bmatrix}$ (red) and $\begin{bmatrix} .7 & .2 \\ .3 & .8 \end{bmatrix}$ (black).

which can be formulated in terms of matrices, related to the singular value decomposition of matrices.

Random matrices are matrices whose entries are random numbers, subject to suitable probability distributions, such as matrix normal distribution. Beyond probability theory, they are applied in domains ranging from number theory to physics.

Symmetries and transformations in physics

Linear transformations and the associated symmetries play a key role in modern physics. For example, elementary particles in quantum field theory are classified as representations of the Lorentz group of special relativity and, more specifically, by their behavior under the spin group. Concrete representations involving the Pauli matrices and more general gamma matrices are an integral part of the physical description of fermions, which behave as spinors. For the three lightest quarks, there is a group-theoretical representation involving the special unitary group $SU(3)$; for their calculations, physicists use a convenient matrix representation known as the Gell-Mann matrices, which are also used for the $SU(3)$ gauge group that forms the basis of the modern description of strong nuclear interactions, quantum chromodynamics. The Cabibbo–Kobayashi–Maskawa matrix, in turn, expresses the fact that the basic quark states that are important for weak interactions are not the same as, but linearly related to the basic quark states that define particles with specific and distinct masses.^[15]

Linear combinations of quantum states

The first model of quantum mechanics (Heisenberg, 1925) represented the theory's operators by infinite-dimensional matrices acting on quantum states. This is also referred to as matrix mechanics. One particular example is the density matrix that characterizes the "mixed" state of a quantum system as a linear combination of elementary, "pure" eigenstates.

Another matrix serves as a key tool for describing the scattering experiments that form the cornerstone of experimental particle physics: Collision reactions such as occur in particle accelerators, where non-interacting particles head towards each other and collide in a small interaction zone, with a new set of non-interacting particles as the result, can be described as the scalar product of outgoing particle states and a linear combination of ingoing particle states. The linear combination is given by a matrix known as the S-matrix, which encodes all information about the possible interactions between particles.

Normal modes

A general application of matrices in physics is to the description of linearly coupled harmonic systems. The equations of motion of such systems can be described in matrix form, with a mass matrix multiplying a generalized velocity to give the kinetic term, and a force matrix multiplying a displacement vector to characterize the interactions. The best way to obtain solutions is to determine the system's eigenvectors, its normal modes, by diagonalizing the matrix equation. Techniques like this are crucial when it comes to the internal dynamics of molecules: the internal vibrations of systems consisting of mutually bound component atoms. They are also needed for describing mechanical vibrations, and oscillations in electrical circuits.

Geometrical optics

Geometrical optics provides further matrix applications. In this approximative theory, the wave nature of light is neglected. The result is a model in which light rays are indeed geometrical rays. If the deflection of light rays by optical elements is small, the action of a lens or reflective element on a given light ray can be expressed as multiplication of a two-component vector with a two-by-two matrix called ray transfer matrix: the vector's components are the light ray's slope and its distance from the optical axis, while the matrix encodes the properties of the optical element. Actually, there are two kinds of matrices, *viz.* a *refraction matrix* describing the refraction at a lens surface, and a *translation matrix*, describing the translation of the plane of reference to the next refracting

surface, where another refraction matrix applies. The optical system, consisting of a combination of lenses and/or reflective elements, is simply described by the matrix resulting from the product of the components' matrices.

Electronics

Traditional mesh analysis in electronics leads to a system of linear equations that can be described with a matrix.

The behaviour of many electronic components can be described using matrices. Let A be a 2-dimensional vector with the component's input voltage v_1 and input current i_1 as its elements, and let B be a 2-dimensional vector with the component's output voltage v_2 and output current i_2 as its elements. Then the behaviour of the electronic component can be described by $B = H \cdot A$, where H is a 2 x 2 matrix containing one impedance element (h_{12}), one admittance element (h_{21}) and two dimensionless elements (h_{11} and h_{22}). Calculating a circuit now reduces to multiplying matrices.

History

Matrices have a long history of application in solving linear equations but they were known as arrays until the 1800s. The Chinese text *The Nine Chapters on the Mathematical Art* is the first example of the use of array methods to solve simultaneous equations,^[16] including the concept of determinants. In 1545 Italian mathematician Girolamo Cardano brought the method to Europe when he published *Ars Magna*.^[17] The Japanese mathematician Seki used the same array methods to solve simultaneous equations in 1683. The Dutch Mathematician Jan de Witt represented transformations using arrays in his 1659 book *Elements of Curves* (1659).^[18] Between 1700 and 1710 Gottfried Wilhelm Leibniz publicized the use of arrays for recording information or solutions and experimented with over 50 different systems of arrays. Cramer presented his rule in 1750.

The term "matrix" (Latin for "womb", derived from *mater*—mother) was coined by James Joseph Sylvester in 1850,^[19] who understood a matrix as an object giving rise to a number of determinants today called minors, that is to say, determinants of smaller matrices that derive from the original one by removing columns and rows. In an 1851 paper, Sylvester explains:

I have in previous papers defined a "Matrix" as a rectangular array of terms, out of which different systems of determinants may be engendered as from the womb of a common parent.^[20]

Arthur Cayley published a treatise on geometric transformations using matrices that were not rotated versions of the coefficients being investigated as had previously been done. Instead he defined operations such as addition, subtraction, multiplication, and division as transformations of those matrices and showed the associative and distributive properties held true. Cayley investigated and demonstrated the non-commutative property of matrix multiplication as well as the commutative property of matrix addition. Early matrix theory had limited the use of arrays almost exclusively to determinants and Arthur Cayley's abstract matrix operations were revolutionary. He was instrumental in proposing a matrix concept independent of equation systems. In 1858 Cayley published his *Memoir on the theory of matrices* in which he proposed and demonstrated the Cayley-Hamilton theorem.

An English mathematician named Cullis was the first to use modern bracket notation for matrices in 1913 and he simultaneously demonstrated the first significant use the notation $\mathbf{A} = [a_{ij}]$ to represent a matrix where a_{ij} refers to the i th row and the j th column.

The study of determinants sprang from several sources. Number-theoretical problems led Gauss to relate coefficients of quadratic forms, i.e., expressions such as $x^2 + xy - 2y^2$, and linear maps in three dimensions to matrices. Eisenstein further developed these notions, including the remark that, in modern parlance, matrix products are non-commutative. Cauchy was the first to prove general statements about determinants, using as definition of the determinant of a matrix $\mathbf{A} = [a_{ij}]$ the following: replace the powers a_j^k by a_{jk} in the polynomial

$$a_1 a_2 \cdots a_n \prod_{i < j} (a_j - a_i),$$

where Π denotes the product of the indicated terms. He also showed, in 1829, that the eigenvalues of symmetric matrices are real. Jacobi studied "functional determinants"—later called Jacobi determinants by Sylvester—which can be used to describe geometric transformations at a local (or infinitesimal) level, see above; Kronecker's *Vorlesungen über die Theorie der Determinanten* and Weierstrass' *Zur Determinantentheorie*, both published in 1903, first treated determinants axiomatically, as opposed to previous more concrete approaches such as the mentioned formula of Cauchy. At that point, determinants were firmly established.

Many theorems were first established for small matrices only, for example the Cayley–Hamilton theorem was proved for 2×2 matrices by Cayley in the aforementioned memoir, and by Hamilton for 4×4 matrices. Frobenius, working on bilinear forms, generalized the theorem to all dimensions (1898). Also at the end of the 19th century the Gauss–Jordan elimination (generalizing a special case now known as Gauss elimination) was established by Jordan. In the early 20th century, matrices attained a central role in linear algebra, partially due to their use in classification of the hypercomplex number systems of the previous century.

The inception of matrix mechanics by Heisenberg, Born and Jordan led to studying matrices with infinitely many rows and columns. Later, von Neumann carried out the mathematical formulation of quantum mechanics, by further developing functional analytic notions such as linear operators on Hilbert spaces, which, very roughly speaking, correspond to Euclidean space, but with an infinity of independent directions.

Other historical usages of the word “matrix” in mathematics

The word has been used in unusual ways by at least two authors of historical importance.

Bertrand Russell and Alfred North Whitehead in their *Principia Mathematica* (1910–1913) use the word “matrix” in the context of their Axiom of reducibility. They proposed this axiom as a means to reduce any function to one of lower type, successively, so that at the “bottom” (0 order) the function is identical to its extension:

“Let us give the name of *matrix* to any function, of however many variables, which does not involve any apparent variables. Then any possible function other than a matrix is derived from a matrix by means of generalization, i.e., by considering the proposition which asserts that the function in question is true with all possible values or with some value of one of the arguments, the other argument or arguments remaining undetermined”.^[21]

For example a function $\Phi(x, y)$ of two variables x and y can be reduced to a *collection* of functions of a single variable, e.g., y , by “considering” the function for all possible values of “individuals” a_i substituted in place of variable x . And then the resulting collection of functions of the single variable y , i.e., $\forall a_i: \Phi(a_i, y)$, can be reduced to a “matrix” of values by “considering” the function for all possible values of “individuals” b_j substituted in place of variable y :

$$\forall b_j \forall a_i: \Phi(a_i, b_j).$$

Alfred Tarski in his 1946 *Introduction to Logic* used the word “matrix” synonymously with the notion of truth table as used in mathematical logic.^[22]

Notes

- [1] equivalently, *table*
- [2] K. Bryan and T. Leise. The \$25,000,000,000 eigenvector: The linear algebra behind Google. *SIAM Review*, 48(3):569–581, 2006.
- [3] <http://ed.ted.com/lessons/how-to-organize-add-and-multiply-matrices-bill-shillito>
- [4] *Eigen* means "own" in German and in Dutch.
- [5] For example, Mathematica, see
- [6] See any standard reference in group.
- [7] Additionally, the group is required to be closed in the general linear group.
- [8] See any reference in representation theory or group representation.
- [9] See the item "Matrix" in
- [10] "Not much of matrix theory carries over to infinite-dimensional spaces, and what does is not so useful, but it sometimes helps."
- [11] "Empty Matrix: A matrix is empty if either its row or column dimension is zero", Glossary (<http://www.omatrix.com/manual/glossary.htm>), O-Matrix v6 User Guide
- [12] "A matrix having at least one dimension equal to zero is called an empty matrix", MATLAB Data Structures (http://www.system.nada.kth.se/unix/software/matlab/Release_14.1/techdoc/matlab_prog/ch_dat29.html)
- [13] . For a more advanced, and more general statement see
- [14] . See also stiffness method.
- [15] see
- [16] cited by
- [17] Discrete Mathematics 4th Ed. Dossey, Otto, Spense, Vanden Eynden, Published by Addison Wesley, October 10, 2001 ISBN 978-0321079121 | p.564-565
- [18] Discrete Mathematics 4th Ed. Dossey, Otto, Spense, Vanden Eynden, Published by Addison Wesley, October 10, 2001 ISBN 978-0321079121 | p.564
- [19] Although many sources state that J. J. Sylvester coined the mathematical term "matrix" in 1848, Sylvester published nothing in 1848. (For proof that Sylvester published nothing in 1848, see: J. J. Sylvester with H. F. Baker, ed., *The Collected Mathematical Papers of James Joseph Sylvester* (Cambridge, England: Cambridge University Press, 1904), vol. 1. (<http://books.google.com/books?id=r-kZAQAIAAJ&pg=PR6#v=onepage&q&f=false>)) His earliest use of the term "matrix" occurs in 1850 in: J. J. Sylvester (1850) "Additions to the articles in the September number of this journal, "On a new class of theorems," and on Pascal's theorem," *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, **37** : 363-370. From page 369 (<http://books.google.com/books?id=CBhDAQAAIAAJ&pg=PA369#v=onepage&q&f=false>): "For this purpose we must commence, not with a square, but with an oblong arrangement of terms consisting, suppose, of m lines and n columns. This will not in itself represent a determinant, but is, as it were, a Matrix out of which we may form various systems of determinants ... "
- [20] The Collected Mathematical Papers of James Joseph Sylvester: 1837–1853, Paper 37 (http://books.google.com/books?id=5GQPlxWrDiEC&pg=PA247&dq=sylvester+matrix+womb&hl=en&ei=uJakTaytCoOv8gPa5cG5Dw&sa=X&oi=book_result&ct=result&resnum=8&ved=0CE8Q6AEwBw#v=onepage&q&f=false), p. 247
- [21] Whitehead, Alfred North; and Russell, Bertrand (1913) *Principia Mathematica to *56*, Cambridge at the University Press, Cambridge UK (republished 1962) cf page 162ff.
- [22] Tarski, Alfred; (1946) *Introduction to Logic and the Methodology of Deductive Sciences*, Dover Publications, Inc, New York NY, ISBN 0-486-28462-X.

References

- Anton, Howard (1987), *Elementary Linear Algebra* (5th ed.), New York: Wiley, ISBN 0-471-84819-0
- Arnold, Vladimir I.; Cooke, Roger (1992), *Ordinary differential equations*, Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-3-540-54813-3
- Artin, Michael (1991), *Algebra*, Prentice Hall, ISBN 978-0-89871-510-1
- Association for Computing Machinery (1979), *Computer Graphics*, Tata McGraw–Hill, ISBN 978-0-07-059376-3
- Baker, Andrew J. (2003), *Matrix Groups: An Introduction to Lie Group Theory*, Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-1-85233-470-3
- Bau III, David; Trefethen, Lloyd N. (1997), *Numerical linear algebra*, Philadelphia, PA: Society for Industrial and Applied Mathematics, ISBN 978-0-89871-361-9
- Beauregard, Raymond A.; Fraleigh, John B. (1973), *A First Course In Linear Algebra: with Optional Introduction to Groups, Rings, and Fields*, Boston: Houghton Mifflin Co., ISBN 0-395-14017-X
- Bretscher, Otto (2005), *Linear Algebra with Applications* (3rd ed.), Prentice Hall

- Bronson, Richard (1989), *Schaum's outline of theory and problems of matrix operations*, New York: McGraw–Hill, ISBN 978-0-07-007978-6
- Brown, William C. (1991), *Matrices and vector spaces*, New York, NY: Marcel Dekker, ISBN 978-0-8247-8419-5
- Coburn, Nathaniel (1955), *Vector and tensor analysis*, New York, NY: Macmillan, OCLC 1029828 (<http://www.worldcat.org/oclc/1029828>)
- Conrey, J. Brian (2007), *Ranks of elliptic curves and random matrix theory*, Cambridge University Press, ISBN 978-0-521-69964-8
- Fraleigh, John B. (1976), *A First Course In Abstract Algebra* (2nd ed.), Reading: Addison-Wesley, ISBN 0-201-01984-1
- Fudenberg, Drew; Tirole, Jean (1983), *Game Theory*, MIT Press
- Gilbarg, David; Trudinger, Neil S. (2001), *Elliptic partial differential equations of second order* (2nd ed.), Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-3-540-41160-4
- Godsil, Chris; Royle, Gordon (2004), *Algebraic Graph Theory*, Graduate Texts in Mathematics **207**, Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-0-387-95220-8
- Golub, Gene H.; Van Loan, Charles F. (1996), *Matrix Computations* (3rd ed.), Johns Hopkins, ISBN 978-0-8018-5414-9
- Greub, Werner Hildbert (1975), *Linear algebra*, Graduate Texts in Mathematics, Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-0-387-90110-7
- Halmos, Paul Richard (1982), *A Hilbert space problem book*, Graduate Texts in Mathematics **19** (2nd ed.), Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-0-387-90685-0, MR 675952 (<http://www.ams.org/mathscinet-getitem?mr=675952>)
- Horn, Roger A.; Johnson, Charles R. (1985), *Matrix Analysis*, Cambridge University Press, ISBN 978-0-521-38632-6
- Householder, Alston S. (1975), *The theory of matrices in numerical analysis*, New York, NY: Dover Publications, MR 0378371 (<http://www.ams.org/mathscinet-getitem?mr=0378371>)
- Krzanowski, Wojtek J. (1988), *Principles of multivariate analysis*, Oxford Statistical Science Series **3**, The Clarendon Press Oxford University Press, ISBN 978-0-19-852211-9, MR 969370 (<http://www.ams.org/mathscinet-getitem?mr=969370>)
- Itō, Kiyosi, ed. (1987), *Encyclopedic dictionary of mathematics. Vol. I-IV* (2nd ed.), MIT Press, ISBN 978-0-262-09026-1, MR 901762 (<http://www.ams.org/mathscinet-getitem?mr=901762>)
- Lang, Serge (1969), *Analysis II*, Addison-Wesley
- Lang, Serge (1987a), *Calculus of several variables* (3rd ed.), Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-0-387-96405-8
- Lang, Serge (1987b), *Linear algebra*, Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-0-387-96412-6
- Lang, Serge (2002), *Algebra*, Graduate Texts in Mathematics **211** (Revised third ed.), New York: Springer-Verlag, ISBN 978-0-387-95385-4, MR 1878556 (<http://www.ams.org/mathscinet-getitem?mr=1878556>)
- Latouche, Guy; Ramaswami, Vaidyanathan (1999), *Introduction to matrix analytic methods in stochastic modeling* (1st ed.), Philadelphia, PA: Society for Industrial and Applied Mathematics, ISBN 978-0-89871-425-8
- Manning, Christopher D.; Schütze, Hinrich (1999), *Foundations of statistical natural language processing*, MIT Press, ISBN 978-0-262-13360-9
- Mehata, K. M.; Srinivasan, S. K. (1978), *Stochastic processes*, New York, NY: McGraw–Hill, ISBN 978-0-07-096612-3
- Mirsky, Leonid (1990), *An Introduction to Linear Algebra* (<http://books.google.com/?id=ULMmheb26ZcC&pg=PA1&dq=linear+algebra+determinant>), Courier Dover Publications, ISBN 978-0-486-66434-7

- Nering, Evar D. (1970), *Linear Algebra and Matrix Theory* (2nd ed.), New York: Wiley, LCCN 76-91646 (<http://lccn.loc.gov/76-91646>)
- Nocedal, Jorge; Wright, Stephen J. (2006), *Numerical Optimization* (2nd ed.), Berlin, DE; New York, NY: Springer-Verlag, p. 449, ISBN 978-0-387-30303-1
- Oualline, Steve (2003), *Practical C++ programming*, O'Reilly, ISBN 978-0-596-00419-4
- Press, William H.; Flannery, Brian P.; Teukolsky, Saul A.; Vetterling, William T. (1992), "LU Decomposition and Its Applications" (http://www.mpi-hd.mpg.de/astrophysik/HEA/internal/Numerical_Recipes/f2-3.pdf), *Numerical Recipes in FORTRAN: The Art of Scientific Computing* (2nd ed.), Cambridge University Press, pp. 34–42
- Punnen, Abraham P.; Gutin, Gregory (2002), *The traveling salesman problem and its variations*, Boston, MA: Kluwer Academic Publishers, ISBN 978-1-4020-0664-7
- Reichl, Linda E. (2004), *The transition to chaos: conservative classical systems and quantum manifestations*, Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-0-387-98788-0
- Rowen, Louis Halle (2008), *Graduate Algebra: noncommutative view*, Providence, RI: American Mathematical Society, ISBN 978-0-8218-4153-2
- Šolin, Pavel (2005), *Partial Differential Equations and the Finite Element Method*, Wiley-Interscience, ISBN 978-0-471-76409-0
- Stinson, Douglas R. (2005), *Cryptography, Discrete Mathematics and its Applications*, Chapman & Hall/CRC, ISBN 978-1-58488-508-5
- Stoer, Josef; Bulirsch, Roland (2002), *Introduction to Numerical Analysis* (3rd ed.), Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-0-387-95452-3
- Ward, J. P. (1997), *Quaternions and Cayley numbers*, Mathematics and its Applications **403**, Dordrecht, NL: Kluwer Academic Publishers Group, ISBN 978-0-7923-4513-8, MR 1458894 (<http://www.ams.org/mathscinet-getitem?mr=1458894>)
- Wolfram, Stephen (2003), *The Mathematica Book* (5th ed.), Champaign, IL: Wolfram Media, ISBN 978-1-57955-022-6

Physics references

- Bohm, Arno (2001), *Quantum Mechanics: Foundations and Applications*, Springer, ISBN 0-387-95330-2
- Burgess, Cliff; Moore, Guy (2007), *The Standard Model. A Primer*, Cambridge University Press, ISBN 0-521-86036-9
- Guenther, Robert D. (1990), *Modern Optics*, John Wiley, ISBN 0-471-60538-7
- Itzykson, Claude; Zuber, Jean-Bernard (1980), *Quantum Field Theory*, McGraw–Hill, ISBN 0-07-032071-3
- Riley, Kenneth F.; Hobson, Michael P.; Bence, Stephen J. (1997), *Mathematical methods for physics and engineering*, Cambridge University Press, ISBN 0-521-55506-X
- Schiff, Leonard I. (1968), *Quantum Mechanics* (3rd ed.), McGraw–Hill
- Weinberg, Steven (1995), *The Quantum Theory of Fields. Volume I: Foundations*, Cambridge University Press, ISBN 0-521-55001-7
- Wherrett, Brian S. (1987), *Group Theory for Atoms, Molecules and Solids*, Prentice–Hall International, ISBN 0-13-365461-3
- Zabrodin, Anton; Brezin, Édouard; Kazakov, Vladimir; Serban, Didina; Wiegmann, Paul (2006), *Applications of Random Matrices in Physics (NATO Science Series II: Mathematics, Physics and Chemistry)*, Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-1-4020-4530-1

Historical references

- Bôcher, Maxime (2004), *Introduction to higher algebra*, New York, NY: Dover Publications, ISBN 978-0-486-49570-5, reprint of the 1907 original edition
- Cayley, Arthur (1889), *The collected mathematical papers of Arthur Cayley* (<http://www.hti.umich.edu/cgi/t/text/pageviewer-idx?c=umhistmath;cc=umhistmath;rgn=full text;idno=ABS3153.0001.001;didno=ABS3153.0001.001;view=image;seq=00000140>), I (1841–1853), Cambridge University Press, pp. 123–126
- Dieudonné, Jean, ed. (1978), *Abrégé d'histoire des mathématiques 1700-1900*, Paris, FR: Hermann
- Hawkins, Thomas (1975), "Cauchy and the spectral theory of matrices", *Historia Mathematica* **2**: 1–29, doi: 10.1016/0315-0860(75)90032-4 ([http://dx.doi.org/10.1016/0315-0860\(75\)90032-4](http://dx.doi.org/10.1016/0315-0860(75)90032-4)), ISSN 0315-0860 (<http://www.worldcat.org/issn/0315-0860>), MR 0469635 (<http://www.ams.org/mathscinet-getitem?mr=0469635>)
- Knobloch, Eberhard (1994), "From Gauss to Weierstrass: determinant theory and its historical evaluations", *The intersection of history and mathematics*, Science Networks Historical Studies **15**, Basel, Boston, Berlin: Birkhäuser, pp. 51–66, MR 1308079 (<http://www.ams.org/mathscinet-getitem?mr=1308079>)
- Kronecker, Leopold (1897), Hensel, Kurt, ed., *Leopold Kronecker's Werke* (<http://name.umdl.umich.edu/AAS8260.0002.001>), Teubner
- Mehra, Jagdish; Rechenberg, Helmut (1987), *The Historical Development of Quantum Theory* (1st ed.), Berlin, DE; New York, NY: Springer-Verlag, ISBN 978-0-387-96284-9
- Shen, Kangshen; Crossley, John N.; Lun, Anthony Wah-Cheung (1999), *Nine Chapters of the Mathematical Art, Companion and Commentary* (2nd ed.), Oxford University Press, ISBN 978-0-19-853936-0
- Weierstrass, Karl (1915), *Collected works* (<http://name.umdl.umich.edu/AAN8481.0003.001>) **3**

External links

Encyclopedic articles

- Hazewinkel, Michiel, ed. (2001), "Matrix" (<http://www.encyclopediaofmath.org/index.php?title=p/m062780>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4

History

- MacTutor: Matrices and determinants (http://www-groups.dcs.st-and.ac.uk/~history/HistTopics/Matrices_and_determinants.html)
- Matrices and Linear Algebra on the Earliest Uses Pages (<http://www.economics.soton.ac.uk/staff/aldrich/matrices.htm>)
- Earliest Uses of Symbols for Matrices and Vectors (<http://jeff560.tripod.com/matrices.html>)

Online books

- Kaw, Autar K., *Introduction to Matrix Algebra* (<http://autarkaw.com/books/matrixalgebra/index.html>), ISBN 978-0-615-25126-4
- *The Matrix Cookbook* (<http://matrixcookbook.com>), retrieved 10 dec 2008
- Brookes, Mike (2005), *The Matrix Reference Manual* (<http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html>), London: Imperial College, retrieved 10 dec 2008

Online matrix calculators

- *SimplyMath (Matrix Calculator)* (<http://www.webalice.it/simoalessia/SimplyMath/matrix.html>)
- *Matrix Calculator (DotNumerics)* (<http://www.dotnumerics.com/MatrixCalculator/>)
- Xiao, Gang, *Matrix calculator* (<http://wims.unice.fr/wims/wims.cgi?module=tool/linear/matrix.en>), retrieved 10 dec 2008
- *Online matrix calculator* (<http://www.bluebit.gr/matrix-calculator/>), retrieved 10 dec 2008
- *Online matrix calculator (ZK framework)* (<http://matrixcalc.info/MatrixZK/>), retrieved 26 nov 2009

- Oehlert, Gary W.; Bingham, Christopher, *MacAnova* (<http://www.stat.umn.edu/macanova/macanova.home.html>), University of Minnesota, School of Statistics, retrieved 10 dec 2008, a freeware package for matrix algebra and statistics
- *Online matrix calculator* (<http://www.idomaths.com/matrix.php>), retrieved 14 dec 2009
- Operation with matrices in R (determinant, track, inverse, adjoint, transpose) (<http://www.elektro-energetika.cz/calculations/matreg.php?language=english>)

Matrix addition

In mathematics, **matrix addition** is the operation of adding two matrices by adding the corresponding entries together. However, there are other operations which could also be considered as a kind of addition for matrices, the direct sum and the Kronecker sum.

Entrywise sum

The usual matrix addition is defined for two matrices of the same dimensions. The sum of two $m \times n$ (pronounced "m by n") matrices **A** and **B**, denoted by **A** + **B**, is again an $m \times n$ matrix computed by adding corresponding elements:^[1]

$$\begin{aligned} \mathbf{A} + \mathbf{B} &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix} \end{aligned}$$

For example:

$$\begin{bmatrix} 1 & 3 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 7 & 5 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1+0 & 3+0 \\ 1+7 & 0+5 \\ 1+2 & 2+1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 8 & 5 \\ 3 & 3 \end{bmatrix}$$

We can also subtract one matrix from another, as long as they have the same dimensions. **A** – **B** is computed by subtracting corresponding elements of **A** and **B**, and has the same dimensions as **A** and **B**. For example:

$$\begin{bmatrix} 1 & 3 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 7 & 5 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1-0 & 3-0 \\ 1-7 & 0-5 \\ 1-2 & 2-1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ -6 & -5 \\ -1 & 1 \end{bmatrix}$$

Direct sum

Another operation, which is used less often, is the direct sum (denoted by \oplus). Note the Kronecker sum is also denoted \oplus ; the context should make the usage clear. The direct sum of any pair of matrices \mathbf{A} of size $m \times n$ and \mathbf{B} of size $p \times q$ is a matrix of size $(m + p) \times (n + q)$ defined as^[1]

$$\mathbf{A} \oplus \mathbf{B} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & b_{11} & \cdots & b_{1q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & b_{p1} & \cdots & b_{pq} \end{bmatrix}$$

For instance,

$$\begin{bmatrix} 1 & 3 & 2 \\ 2 & 3 & 1 \end{bmatrix} \oplus \begin{bmatrix} 1 & 6 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 2 & 0 & 0 \\ 2 & 3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The direct sum of matrices is a special type of block matrix, in particular the direct sum of square matrices is a block diagonal matrix.

The adjacency matrix of the union of disjoint graphs or multigraphs is the direct sum of their adjacency matrices. Any element in the direct sum of two vector spaces of matrices can be represented as a direct sum of two matrices.

In general, the direct sum of n matrices is:^[1]

$$\bigoplus_{i=1}^n \mathbf{A}_i = \text{diag}(\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3 \cdots \mathbf{A}_n) = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_n \end{bmatrix}$$

where the zeros are actually blocks of zeros, i.e. zero matrices.

NB: Sometimes in this context, boldtype for matrices is dropped, matrices are written in italic.

Kronecker sum

The Kronecker sum is different from the direct sum but is also denoted by \oplus . It is defined using the Kronecker product \otimes and normal matrix addition. If \mathbf{A} is n -by- n , \mathbf{B} is m -by- m and \mathbf{I}_k denotes the k -by- k identity matrix then the Kronecker sum is defined by:

$$\mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I}_m + \mathbf{I}_n \otimes \mathbf{B}.$$

Notes

[1] Lipschutz Lipson.

References

- Lipschutz, S.; Lipson, M. (2009). *Linear Algebra*. Schaum's Outline Series. ISBN 978-0-07-154352-1.

External links

- Direct sum of matrices (<http://planetmath.org/encyclopedia/DirectSumOfMatrices.html>) at PlanetMath
- 4x4 Matrix Addition and Subtraction (<http://ncalculators.com/matrix/4x4-matrix-addition-subtraction-calculator.htm>)
- Abstract nonsense: Direct Sum of Linear Transformations and Direct Sum of Matrices (<http://drexel28.wordpress.com/2010/12/22/direct-sum-of-linear-transformations-and-direct-sum-of-matrices-pt-iii/>)
- Mathematics Source Library: Arithmetic Matrix Operations (http://www.mymathlib.com/matrices/arithmetic/direct_sum.html)
- Matrix Algebra and R (<http://www.aps.uoguelph.ca/~lrs/ABMethods/NOTES/CDmatrix.pdf>)

Matrix multiplication

In mathematics, **matrix multiplication** is a binary operation that takes a pair of matrices, and produces another matrix. Numbers such as the real or complex numbers can be multiplied according to elementary arithmetic. On the other hand, matrices are *arrays of numbers*, so there is no unique way to define "the" multiplication of matrices. As such, in general the term "matrix multiplication" refers to a number of different ways to multiply matrices. The key features of any matrix multiplication include: the number of rows and columns the original matrices have (called the "size", "order" or "dimension"), and specifying how the entries of the matrices generate the new matrix.

Like vectors, matrices of any size can be multiplied by scalars, which amounts to multiplying every entry of the matrix by the same number. Similar to the entrywise definition of adding or subtracting matrices, multiplication of two matrices of the same size can be defined by multiplying the corresponding entries, and this is known as the Hadamard product. Another definition is the Kronecker product of two matrices, to obtain a block matrix.

One can form many other definitions. However, the most useful definition can be motivated by linear equations and linear transformations on vectors, which have numerous applications in applied mathematics, physics, and engineering. This definition is often called *the matrix product*.^{[1][2]} In words, if **A** is an $n \times m$ matrix and **B** is a $m \times p$ matrix, their matrix product **AB** is an $n \times p$ matrix, in which the m entries across the rows of **A** are multiplied with the m entries down the columns of **B** (the precise definition is below).

This definition is not commutative, although it still retains the associative property and is distributive over entrywise addition of matrices. The identity element of the matrix product is the identity matrix (analogous to multiplying numbers by 1), and a square matrix may have an inverse matrix (analogous to the multiplicative inverse of a number). A consequence of the matrix product is determinant multiplicativity. The matrix product is an important operation in linear transformations, matrix groups, and the theory of group representations and irreps. For large matrices and/or products of more than two matrices, this matrix product can be very time consuming to calculate, so more efficient algorithms to compute the matrix product than the mathematical definition have been developed.

This article will use the following notational conventions: matrices are represented by capital letters in bold, e.g. **A**, vectors in lowercase bold, e.g. **a**, and entries of vectors and matrices are italic (since they are scalars), e.g. *A* and *a*. Index notation is often the clearest way to express definitions, and will be used as standard in the literature. The i, j entry of matrix **A** is indicated by $(\mathbf{A})_{ij}$ or A_{ij} , whereas a numerical label (not matrix entries) on a collection of matrices is subscripted only, e.g. $\mathbf{A}_1, \mathbf{A}_2$, etc.

Scalar multiplication

The simplest form of multiplication associated with matrices is scalar multiplication.

The **left scalar multiplication** of a matrix \mathbf{A} with a scalar λ gives another matrix $\lambda\mathbf{A}$ of the same size as \mathbf{A} . The entries of $\lambda\mathbf{A}$ are defined by

$$(\lambda\mathbf{A})_{ij} = \lambda(\mathbf{A})_{ij},$$

explicitly:

$$\lambda\mathbf{A} = \lambda \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} = \begin{pmatrix} \lambda A_{11} & \lambda A_{12} & \cdots & \lambda A_{1m} \\ \lambda A_{21} & \lambda A_{22} & \cdots & \lambda A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda A_{n1} & \lambda A_{n2} & \cdots & \lambda A_{nm} \end{pmatrix}.$$

Similarly, the **right scalar multiplication** of a matrix \mathbf{A} with a scalar λ is defined to be

$$(\mathbf{A}\lambda)_{ij} = (\mathbf{A})_{ij}\lambda,$$

explicitly:

$$\mathbf{A}\lambda = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} \lambda = \begin{pmatrix} A_{11}\lambda & A_{12}\lambda & \cdots & A_{1m}\lambda \\ A_{21}\lambda & A_{22}\lambda & \cdots & A_{2m}\lambda \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1}\lambda & A_{n2}\lambda & \cdots & A_{nm}\lambda \end{pmatrix}.$$

When the underlying ring is commutative, for example, the real or complex number field, these two multiplications are the same, and are simply called *scalar multiplication*. However, for matrices over a more general ring that are *not* commutative, such as the quaternions, they may not be equal.

For a real scalar and matrix:

$$\lambda = 2, \quad \mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$2\mathbf{A} = 2 \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 2 \cdot a & 2 \cdot b \\ 2 \cdot c & 2 \cdot d \end{pmatrix} = \begin{pmatrix} a \cdot 2 & b \cdot 2 \\ c \cdot 2 & d \cdot 2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} 2 = \mathbf{A}2.$$

For quaternion scalars and matrices:

$$\lambda = i, \quad \mathbf{A} = \begin{pmatrix} i & 0 \\ 0 & j \end{pmatrix}$$

$$i \begin{pmatrix} i & 0 \\ 0 & j \end{pmatrix} = \begin{pmatrix} i^2 & 0 \\ 0 & ij \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & k \end{pmatrix} \neq \begin{pmatrix} -1 & 0 \\ 0 & -k \end{pmatrix} = \begin{pmatrix} i^2 & 0 \\ 0 & ji \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & j \end{pmatrix} i,$$

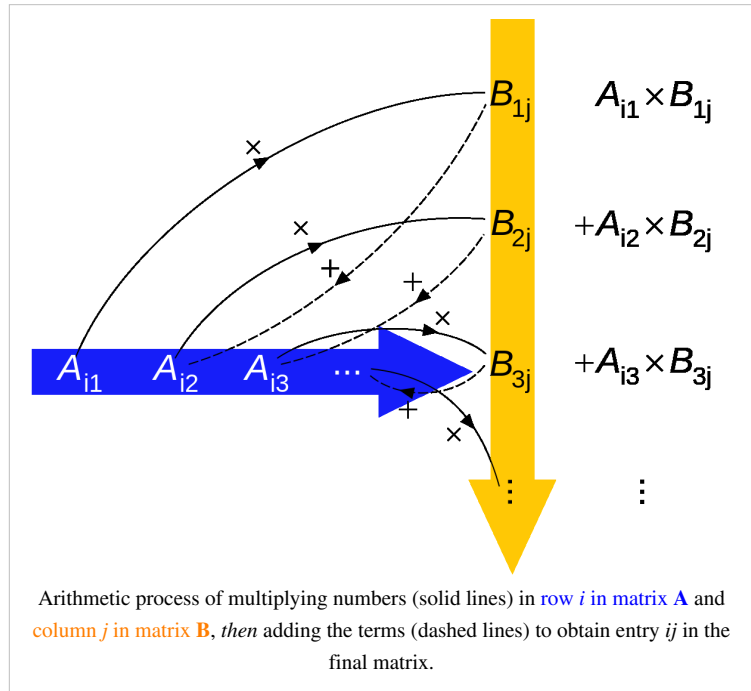
where i, j, k are the quaternion units. The non-commutativity of quaternion multiplication prevents the transition of changing $ij = +k$ to $ji = -k$.

Matrix product (two matrices)

Assume two matrices are to be multiplied (the generalization to any number is discussed below).

General definition of the matrix product

If \mathbf{A} is an $n \times m$ matrix and \mathbf{B} is an $m \times p$ matrix,



$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mp} \end{pmatrix}$$

the **matrix product** \mathbf{AB} (denoted without multiplication signs or dots) is defined to be the $n \times p$ matrix^{[3][4]}

$$\mathbf{AB} = \begin{pmatrix} (\mathbf{AB})_{11} & (\mathbf{AB})_{12} & \cdots & (\mathbf{AB})_{1p} \\ (\mathbf{AB})_{21} & (\mathbf{AB})_{22} & \cdots & (\mathbf{AB})_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{AB})_{n1} & (\mathbf{AB})_{n2} & \cdots & (\mathbf{AB})_{np} \end{pmatrix}$$

where each i, j entry is given by multiplying the entries A_{ik} (across row i of \mathbf{A}) by the entries B_{kj} (down column j of \mathbf{B}), for $k = 1, 2, \dots, m$, and summing the results over k :

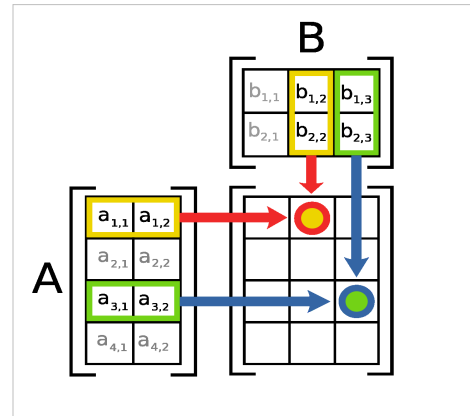
$$(\mathbf{AB})_{ij} = \sum_{k=1}^m A_{ik} B_{kj}.$$

Thus the product \mathbf{AB} is defined only if the number of columns in \mathbf{A} is equal to the number of rows in \mathbf{B} , in this case m . Each entry may be computed one at a time. Sometimes, the summation convention is used as it is understood to sum over the repeated index k . To prevent any ambiguity, this convention will not be used in the article.

Usually the entries are numbers or expressions, but can even be matrices themselves (see block matrix). The matrix product can still be calculated exactly the same way. See below for details on how the matrix product can be calculated in terms of blocks taking the forms of rows and columns.

Illustration

The figure to the right illustrates diagrammatically the product of two matrices **A** and **B**, showing how each intersection in the product matrix corresponds to a row of **A** and a column of **B**.



$$\begin{array}{c} 4 \times 2 \text{ matrix} \\ \begin{bmatrix} a_{1,1} & a_{1,2} \\ \cdot & \cdot \\ a_{3,1} & a_{3,2} \\ \cdot & \cdot \end{bmatrix} \end{array} \begin{array}{c} 2 \times 3 \text{ matrix} \\ \begin{bmatrix} \cdot & b_{1,2} & b_{1,3} \\ \cdot & b_{2,2} & b_{2,3} \end{bmatrix} \end{array} = \begin{array}{c} 4 \times 3 \text{ matrix} \\ \begin{bmatrix} \cdot & x_{12} & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & x_{33} \\ \cdot & \cdot & \cdot \end{bmatrix} \end{array}$$

The values at the intersections marked with circles are:

$$x_{12} = a_{11}b_{12} + a_{12}b_{22}$$

$$x_{33} = a_{31}b_{13} + a_{32}b_{23}$$

Examples of matrix products

Row vector and column vector

If

$$\mathbf{A} = (a \ b), \quad \mathbf{B} = \begin{pmatrix} x \\ y \end{pmatrix},$$

their matrix products are:

$$\mathbf{AB} = (a \ b) \begin{pmatrix} x \\ y \end{pmatrix} = ax + by,$$

and

$$\mathbf{BA} = \begin{pmatrix} x \\ y \end{pmatrix} (a \ b) = \begin{pmatrix} xa & xb \\ ya & yb \end{pmatrix}.$$

Note **AB** and **BA** are two very different matrices: the first is a 1×1 matrix while the second is a 2×2 matrix. Such expressions occur for real-valued Euclidean vectors in Cartesian coordinates, displayed as row and column matrices, in which case **AB** is the matrix form of their inner product, while **BA** the matrix form of their dyadic or tensor product.

Square matrix and column vector

If

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} x \\ y \end{pmatrix},$$

their matrix product is:

$$\mathbf{AB} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix},$$

however \mathbf{BA} is not defined.

The product of a square matrix multiplied by a column matrix arises naturally in linear algebra; for solving linear equations and representing linear transformations. By choosing a, b, c, d in \mathbf{A} appropriately, \mathbf{A} can represent a variety of transformations such as rotations, scaling and reflections, shears, of a geometric shape in space.

Square matrices

If

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix},$$

their matrix products are:

$$\mathbf{AB} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} a\alpha + b\gamma & a\beta + b\delta \\ c\alpha + d\gamma & c\beta + d\delta \end{pmatrix},$$

and

$$\mathbf{BA} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \alpha a + \beta c & \alpha b + \beta d \\ \gamma a + \delta c & \gamma b + \delta d \end{pmatrix}.$$

In this case, both products \mathbf{AB} and \mathbf{BA} are defined, and the entries show that \mathbf{AB} and \mathbf{BA} are not equal in general. Multiplying square matrices which represent linear transformations corresponds to the composite transformation (see below for details).

Row vector, square matrix, and column vector

If

$$\mathbf{A} = (a \ b), \quad \mathbf{B} = \begin{pmatrix} p & q \\ r & s \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} x \\ y \end{pmatrix},$$

their matrix product is:

$$\begin{aligned} \mathbf{ABC} &= (a \ b) \left[\begin{pmatrix} p & q \\ r & s \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right] = \left[(a \ b) \begin{pmatrix} p & q \\ r & s \end{pmatrix} \right] \begin{pmatrix} x \\ y \end{pmatrix} \\ &= (a \ b) \begin{pmatrix} px + qy \\ rx + sy \end{pmatrix} = (ap + br \quad aq + bs) \begin{pmatrix} x \\ y \end{pmatrix} \\ &= apx + aqy + brx + bsy, \end{aligned}$$

however \mathbf{CBA} is not defined. Note that $\mathbf{A(BC)} = (\mathbf{AB})\mathbf{C}$, this is one of many general properties listed below. Expressions of the form \mathbf{ABC} occur when calculating the inner product of two vectors displayed as row and column vectors in an arbitrary coordinate system, and the metric tensor in these coordinates written as the square matrix.

Properties of the matrix product (two matrices)

Analogous to numbers (elements of a field), matrices satisfy the following general properties, although there is one subtlety, due to the nature of matrix multiplication.

All matrices

1. Not commutative:

In general:

$$\mathbf{AB} \neq \mathbf{BA}$$

because \mathbf{AB} and \mathbf{BA} may not be simultaneously defined, and even if they are they may still not be equal. This is contrary to ordinary multiplication of numbers. To specify the ordering of matrix multiplication in words; "pre-multiply (or left multiply) \mathbf{A} by \mathbf{B} " means \mathbf{BA} , while "post-multiply (or right multiply) \mathbf{A} by \mathbf{C} " means \mathbf{AC} . As long as the entries of the matrix come from a ring that has an identity, and $n > 1$ there is a pair of $n \times n$

noncommuting matrices over the ring. A notable exception is that the identity matrix (or any scalar multiple of it) commutes with every square matrix.

In index notation:

$$\sum_k A_{ik} B_{kj} \neq \sum_k B_{ik} A_{kj}$$

2. **Distributive over matrix addition:**

Left distributivity:

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$$

Right distributivity:

$$(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$$

In index notation, these are respectively:

$$\begin{aligned} \sum_k A_{ik}(B_{kj} + C_{kj}) &= \sum_k A_{ik}B_{kj} + \sum_k A_{ik}C_{kj} \\ \sum_k (A_{ik} + B_{ik})C_{kj} &= \sum_k A_{ik}C_{kj} + \sum_k B_{ik}C_{kj} \end{aligned}$$

3. **Scalar multiplication is compatible with matrix multiplication:**

$$\lambda(\mathbf{AB}) = (\lambda\mathbf{A})\mathbf{B} \text{ and } (\mathbf{AB})\lambda = \mathbf{A}(\mathbf{B}\lambda)$$

where λ is a scalar. If the entries of the matrix are real or complex numbers (or from any other commutative ring), then all four quantities are equal. More generally, all four are equal if λ belongs to the center of the ring of entries of the matrix, because in this case $\lambda\mathbf{X} = \mathbf{X}\lambda$ for all matrices \mathbf{X} .

In index notation, these are respectively:

$$\begin{aligned} \lambda \sum_k (A_{ik} B_{kj}) &= \sum_k (\lambda A_{ik}) B_{kj} = \sum_k A_{ik} (\lambda B_{kj}) \\ \sum_k (A_{ik} B_{kj}) \lambda &= \sum_k (A_{ik} \lambda) B_{kj} = \sum_k A_{ik} (B_{kj} \lambda) \end{aligned}$$

4. **Transpose:**

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$$

where T denotes the transpose, the interchange of row i with column i in a matrix. This identity holds for any matrices over a commutative ring, but not for all rings in general. Note that 'A and \mathbf{B} are reversed.

In index notation:

$$[(\mathbf{AB})^T]_{ij} = (\mathbf{AB})_{ji} = \sum_k (\mathbf{A})_{jk} (\mathbf{B})_{ki} = \sum_k (\mathbf{A}^T)_{kj} (\mathbf{B}^T)_{ik} = \sum_k (\mathbf{B}^T)_{ik} (\mathbf{A}^T)_{kj} = [(\mathbf{A}^T) (\mathbf{B}^T)]_{ij}$$

5. **Complex conjugate:** If \mathbf{A} and \mathbf{B} have complex entries, then

$$(\mathbf{AB})^* = \mathbf{A}^* \mathbf{B}^*$$

where $*$ denotes the complex conjugate of a matrix.

In index notation:

$$[(\mathbf{AB})^*]_{ij} = \left[\sum_k (\mathbf{A})_{ik} (\mathbf{B})_{kj} \right]^* = \sum_k (\mathbf{A})_{ik}^* (\mathbf{B})_{kj}^* = \sum_k (\mathbf{A}^*)_{ik} (\mathbf{B}^*)_{kj} = (\mathbf{A}^* \mathbf{B}^*)_{ij}$$

6. **Conjugate transpose:** If \mathbf{A} and \mathbf{B} have complex entries, then

$$(\mathbf{AB})^\dagger = \mathbf{B}^\dagger \mathbf{A}^\dagger$$

where \dagger denotes the Conjugate transpose of a matrix (complex conjugate and transposed).

In index notation:

$$[(\mathbf{AB})^\dagger]_{ij} = [(\mathbf{AB})^*]_{ji} = \sum_k (\mathbf{A}^*)_{jk} (\mathbf{B}^*)_{ki} = \sum_k (\mathbf{A}^\dagger)_{kj} (\mathbf{B}^\dagger)_{ik} = \sum_k (\mathbf{B}^\dagger)_{ik} (\mathbf{A}^\dagger)_{kj} = [(\mathbf{A}^\dagger) (\mathbf{B}^\dagger)]_{ij}$$

7. **Traces:** The trace of a product \mathbf{AB} is independent of the order of \mathbf{A} and \mathbf{B} :

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$$

In index notation:

$$\text{tr}(\mathbf{AB}) = \sum_i \sum_k A_{ik} B_{ki} = \sum_k \sum_i B_{ki} A_{ik} = \text{tr}(\mathbf{BA})$$

Square matrices only

1. **Identity element:** If \mathbf{A} is a square matrix, then

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

where \mathbf{I} is the identity matrix of the same order.

2. **Inverse matrix:** If \mathbf{A} is a square matrix, there *may* be an inverse matrix \mathbf{A}^{-1} of \mathbf{A} such that

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

If this property holds then \mathbf{A} is an invertible matrix, if not \mathbf{A} is a singular matrix. Moreover,

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

3. **Determinants:** The determinant of a product \mathbf{AB} is the product of the determinants of square matrices \mathbf{A} and \mathbf{B} (not defined when the underlying ring is not commutative):

$$\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$$

Since $\det(\mathbf{A})$ and $\det(\mathbf{B})$ are just numbers and so commute, $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B}) = \det(\mathbf{B})\det(\mathbf{A}) = \det(\mathbf{BA})$, even when $\mathbf{AB} \neq \mathbf{BA}$.

Matrix product (any number)

Matrix multiplication can be extended to the case of more than two matrices, provided that for each sequential pair, their dimensions match.

The product of n matrices $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_n$ with sizes $s_0 \times s_1, s_1 \times s_2, \dots, s_{n-1} \times s_n$ (where $s_0, s_1, s_2, \dots, s_n$ are all simply positive integers and the subscripts are labels corresponding to the matrices, nothing more), is the $s_0 \times s_n$ matrix:

$$\prod_{i=1}^n \mathbf{A}_i = \mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_n.$$

In index notation:

$$(\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_n)_{i_0 i_n} = \sum_{i_1=1}^{s_1} \sum_{i_2=1}^{s_2} \cdots \sum_{i_{n-1}=1}^{s_{n-1}} (\mathbf{A}_1)_{i_0 i_1} (\mathbf{A}_2)_{i_1 i_2} (\mathbf{A}_3)_{i_2 i_3} \cdots (\mathbf{A}_{n-1})_{i_{n-2} i_{n-1}} (\mathbf{A}_n)_{i_{n-1} i_n}$$

Properties of the matrix product (any number)

The same properties will hold, as long as the ordering of matrices is not changed. Some of the previous properties for more than two matrices generalize as follows.

1. **Associative:**

The matrix product is associative. If three matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} are respectively $m \times p$, $p \times q$, and $q \times r$ matrices, then there are two ways of grouping them without changing their order, and

$$\mathbf{ABC} = \mathbf{A(BC)} = (\mathbf{AB})\mathbf{C}$$

is an $m \times r$ matrix.

If four matrices \mathbf{A} , \mathbf{B} , \mathbf{C} , and \mathbf{D} are respectively $m \times p$, $p \times q$, $q \times r$, and $r \times s$ matrices, then there are five ways of grouping them without changing their order, and

$$\mathbf{ABCD} = ((\mathbf{AB})\mathbf{C})\mathbf{D} = (\mathbf{A(BC)})\mathbf{D} = \mathbf{A((BC)\mathbf{D})} = \mathbf{A(B(CD))} = (\mathbf{AB})(\mathbf{CD})$$

is an $m \times s$ matrix.

In general, the number of possible ways of grouping n matrices for multiplication is equal to the $(n - 1)$ th Catalan number

2. **Trace:** The trace of a product of n matrices $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_n$ is invariant under cyclic permutations of the matrices in the product:

$$\text{tr}(\mathbf{A}_1 \mathbf{A}_2 \mathbf{A}_3 \dots \mathbf{A}_{n-2} \mathbf{A}_{n-1} \mathbf{A}_n) = \text{tr}(\mathbf{A}_2 \mathbf{A}_3 \mathbf{A}_4 \dots \mathbf{A}_{n-1} \mathbf{A}_n \mathbf{A}_1) = \text{tr}(\mathbf{A}_3 \mathbf{A}_4 \mathbf{A}_5 \dots \mathbf{A}_n \mathbf{A}_1 \mathbf{A}_2) = \dots$$

3. **Determinant:** For square matrices only, the determinant of a product is the product of determinants:

$$\det \left(\prod_{i=1}^n \mathbf{A}_i \right) = \prod_{i=1}^n \det(\mathbf{A}_i)$$

Examples of chain multiplication

Similarity transformations involving similar matrices are matrix products of the three square matrices, in the form:

$$\mathbf{B} = \mathbf{P}^{-1} \mathbf{A} \mathbf{P}$$

where \mathbf{P} is the similarity matrix and \mathbf{A} and \mathbf{B} are said to be similar if this relation holds. This product appears frequently in linear algebra and applications, such as diagonalizing square matrices and the equivalence between different matrix representations of the same linear operator.

Operations derived from the matrix product

More operations on square matrices can be defined using the matrix product, such as powers and n th roots by repeated matrix products, the matrix exponential can be defined by a power series, the matrix logarithm is the inverse of matrix exponentiation, and so on.

Powers of matrices

Square matrices can be multiplied by themselves repeatedly in the same way as ordinary numbers, because they always have the same number of rows and columns. This repeated multiplication can be described as a **power of the matrix**, a special case of the ordinary matrix product. On the contrary, *rectangular* matrices do not have the same number of rows and columns so they can *never* be raised to a power. An $n \times n$ matrix \mathbf{A} raised to a positive integer k is defined as

$$\mathbf{A}^k = \underbrace{\mathbf{A} \mathbf{A} \dots \mathbf{A}}_{k \text{ times}}$$

and the following identities hold, where λ is a scalar:

1. **Zero power:**

$$\mathbf{A}^0 = \mathbf{I}$$

where \mathbf{I} is the identity matrix. This is parallel to the zeroth power of any number which equals unity.

2. **Scalar multiplication:**

$$(\lambda \mathbf{A})^k = \lambda^k \mathbf{A}^k$$

3. **Determinant:**

$$\det(\mathbf{A}^k) = \det(\mathbf{A})^k$$

The naive computation of matrix powers is to multiply k times the matrix \mathbf{A} to the result, starting with the identity matrix just like the scalar case. This can be improved using exponentiation by squaring, a method commonly used for scalars. For diagonalizable matrices, an even better method is to use the eigenvalue decomposition of \mathbf{A} . Another method based on the Cayley–Hamilton theorem finds an identity using the matrices' characteristic polynomial, producing a more effective equation for \mathbf{A}^k in which a *scalar* is raised to the required power, rather than an entire *matrix*.

A special case is the power of a diagonal matrix. Since the product of diagonal matrices amounts to simply multiplying corresponding diagonal elements together, the power k of a diagonal matrix \mathbf{A} will have entries raised to the power. Explicitly;

$$\mathbf{A}^k = \begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{pmatrix}^k = \begin{pmatrix} A_{11}^k & 0 & \cdots & 0 \\ 0 & A_{22}^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn}^k \end{pmatrix}$$

meaning it is easy to raise a diagonal matrix to a power. When raising an arbitrary matrix (not necessarily a diagonal matrix) to a power, it is often helpful to exploit this property by diagonalizing the matrix first.

Applications of the matrix product

Linear transformations

Matrices offer a concise way of representing linear transformations between vector spaces, and matrix multiplication corresponds to the composition of linear transformations. The matrix product of two matrices can be defined when their entries belong to the same ring, and hence can be added and multiplied.

Let U , V , and W be vector spaces over the same field with given bases, $S: V \rightarrow W$ and $T: U \rightarrow V$ be linear transformations and $ST: U \rightarrow W$ be their composition.

Suppose that \mathbf{A} , \mathbf{B} , and \mathbf{C} are the matrices representing the transformations S , T , and ST with respect to the given bases.

Then $\mathbf{AB} = \mathbf{C}$, that is, the matrix of the composition (or the product) of linear transformations is the product of their matrices with respect to the given bases.

Linear systems of equations

A system of linear equations can be solved by collecting the coefficients of the equations into a square matrix, then inverting the matrix equation.

A similar procedure can be used to solve a system of linear differential equations, see also phase plane.

The inner and outer products

Given two *column vectors* \mathbf{a} and \mathbf{b} , the Euclidean inner product and outer product are the simplest special cases of the matrix product, by transposing the column vectors into row vectors.^[5]

Inner product

The **inner product** of two vectors in matrix form is equivalent to a column vector multiplied on the left by a row vector:

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b} = (a_1 \ a_2 \ \cdots \ a_n) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{i=1}^n a_i b_i$$

The matrix product itself can be expressed in terms of inner product. Suppose that the first $n \times m$ matrix \mathbf{A} is decomposed into its row vectors \mathbf{a}_i , and the second $m \times p$ matrix \mathbf{B} into its column vectors \mathbf{b}_i :

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mp} \end{pmatrix} = (\mathbf{b}_1 \ \mathbf{b}_2 \ \cdots \ \mathbf{b}_p)$$

where

$$\mathbf{a}_i = (A_{i1} \ A_{i2} \ \cdots \ A_{im}), \quad \mathbf{b}_i = \begin{pmatrix} B_{1i} \\ B_{2i} \\ \vdots \\ B_{mi} \end{pmatrix}$$

Then:

$$\mathbf{AB} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} (\mathbf{b}_1 \ \mathbf{b}_2 \ \cdots \ \mathbf{b}_p) = \begin{pmatrix} (\mathbf{a}_1 \cdot \mathbf{b}_1) & (\mathbf{a}_1 \cdot \mathbf{b}_2) & \cdots & (\mathbf{a}_1 \cdot \mathbf{b}_p) \\ (\mathbf{a}_2 \cdot \mathbf{b}_1) & (\mathbf{a}_2 \cdot \mathbf{b}_2) & \cdots & (\mathbf{a}_2 \cdot \mathbf{b}_p) \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{a}_n \cdot \mathbf{b}_1) & (\mathbf{a}_n \cdot \mathbf{b}_2) & \cdots & (\mathbf{a}_n \cdot \mathbf{b}_p) \end{pmatrix}$$

It is also possible to express a matrix product in terms of concatenations of products of matrices and row or column vectors:

$$\mathbf{AB} = (\mathbf{Ab}_1 \ \mathbf{Ab}_2 \ \cdots \ \mathbf{Ab}_p) = \begin{pmatrix} \mathbf{a}_1\mathbf{B} \\ \mathbf{a}_2\mathbf{B} \\ \vdots \\ \mathbf{a}_n\mathbf{B} \end{pmatrix}$$

These decompositions are particularly useful for matrices that are envisioned as concatenations of particular types of row vectors or column vectors, e.g. orthogonal matrices (whose rows and columns are unit vectors orthogonal to each other) and Markov matrices (whose rows or columns sum to 1).

Outer product

The **outer product** (also known as the **dyadic product** or **tensor product**) of two vectors in matrix form is equivalent to a row vector multiplied on the left by a column vector:

$$\mathbf{a} \otimes \mathbf{b} = \mathbf{ab}^T = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} (b_1 \ b_2 \ \cdots \ b_n) = \begin{pmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{pmatrix}.$$

An alternative method is to express the matrix product in terms of the outer product. The decomposition is done the other way around, the first matrix \mathbf{A} is decomposed into column vectors \mathbf{a}_i and the second matrix \mathbf{B} into row vectors $\bar{\mathbf{b}}_i$:

$$\mathbf{AB} = (\bar{\mathbf{a}}_1 \ \bar{\mathbf{a}}_2 \ \cdots \ \bar{\mathbf{a}}_m) \begin{pmatrix} \bar{\mathbf{b}}_1 \\ \bar{\mathbf{b}}_2 \\ \vdots \\ \bar{\mathbf{b}}_m \end{pmatrix} = \bar{\mathbf{a}}_1 \otimes \bar{\mathbf{b}}_1 + \bar{\mathbf{a}}_2 \otimes \bar{\mathbf{b}}_2 + \cdots + \bar{\mathbf{a}}_m \otimes \bar{\mathbf{b}}_m = \sum_{i=1}^m \bar{\mathbf{a}}_i \otimes \bar{\mathbf{b}}_i$$

where this time

$$\bar{\mathbf{a}}_i = \begin{pmatrix} A_{1i} \\ A_{2i} \\ \vdots \\ A_{ni} \end{pmatrix}, \quad \bar{\mathbf{b}}_i = (B_{i1} \ B_{i2} \ \cdots \ B_{ip}).$$

This method emphasizes the effect of individual column/row pairs on the result, which is a useful point of view with e.g. covariance matrices, where each such pair corresponds to the effect of a single sample point.

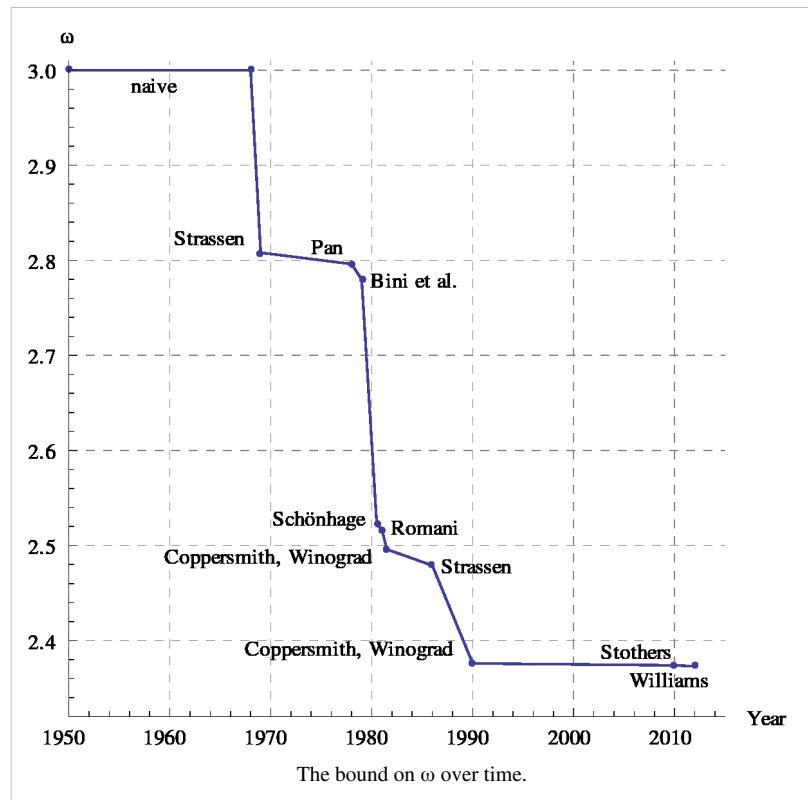
$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \\ 7 \end{pmatrix} (a \ d) + \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} (b \ e) + \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix} (c \ f) = \begin{pmatrix} 1a & 1d \\ 4a & 4d \\ 7a & 7d \end{pmatrix} + \begin{pmatrix} 2b & 2e \\ 5b & 5e \\ 8b & 8e \end{pmatrix} + \begin{pmatrix} 3c & 3f \\ 6c & 6f \\ 9c & 9f \end{pmatrix}.$$

Algorithms for efficient matrix multiplication

List of unsolved problems in computer science

What is the fastest algorithm for matrix multiplication?

The running time of square matrix multiplication, if carried out naïvely, is $O(n^3)$. The running time for multiplying rectangular matrices (one $m \times p$ -matrix with one $p \times n$ -matrix) is $O(mnp)$, however, more efficient algorithms exist, such as Strassen's algorithm, devised by Volker Strassen in 1969 and often referred to as "fast matrix multiplication". It is based on a way of multiplying two 2×2 -matrices which requires only 7 multiplications (instead of the usual 8), at the expense of several additional addition and subtraction operations. Applying this recursively gives an algorithm with a multiplicative cost of $O(n^{\log_2 7}) \approx O(n^{2.807})$. Strassen's algorithm is more complex, and the numerical stability is reduced compared to the naïve algorithm.



Nevertheless, it appears in several libraries, such as BLAS, where it is significantly more efficient for matrices with dimensions $n > 100$,^[6] and is very useful for large matrices over exact domains such as finite fields, where numerical stability is not an issue.

The current $O(n^k)$ algorithm with the lowest known exponent k is a generalization of the Coppersmith–Winograd algorithm that has an asymptotic complexity of $O(n^{2.3729})$ thanks to Vassilevska Williams.^[7] This algorithm, and the Coppersmith–Winograd algorithm on which it is based, are similar to Strassen's algorithm: a way is devised for multiplying two $k \times k$ -matrices with fewer than k^3 multiplications, and this technique is applied recursively. However, the constant coefficient hidden by the Big O notation is so large that these algorithms are only worthwhile for matrices that are too large to handle on present-day computers.

Since any algorithm for multiplying two $n \times n$ -matrices has to process all $2 \times n^2$ -entries, there is an asymptotic lower bound of $\Omega(n^2)$ operations. Raz (2002) proves a lower bound of $\Omega(n^2 \log(n))$ for bounded coefficient arithmetic circuits over the real or complex numbers.

Cohn *et al.* (2003, 2005) put methods such as the Strassen and Coppersmith–Winograd algorithms in an entirely different group-theoretic context, by utilising triples of subsets of finite groups which satisfy a disjointness property called the triple product property (TPP). They show that if families of wreath products of Abelian groups with symmetric groups realise families of subset triples with a simultaneous version of the TPP, then there are matrix multiplication algorithms with essentially quadratic complexity. Most researchers believe that this is indeed the case.^[8] However, Alon, Shpilka and Umans have recently shown that some of these conjectures implying fast matrix multiplication are incompatible with another plausible conjecture, the sunflower conjecture.^[9]

Because of the nature of matrix operations and the layout of matrices in memory, it is typically possible to gain substantial performance gains through use of parallelization and vectorization. It should therefore be noted that some lower time-complexity algorithms on paper may have indirect time complexity costs on real machines.

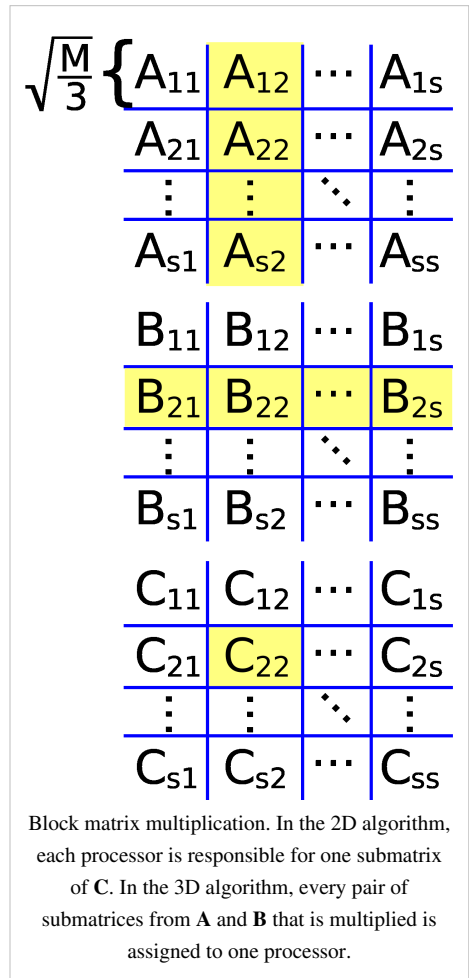
Freivalds' algorithm is a simple Monte Carlo algorithm that given matrices \mathbf{A} , \mathbf{B} , \mathbf{C} verifies in $\Theta(n^2)$ time if $\mathbf{AB} = \mathbf{C}$.

Communication-avoiding and distributed algorithms

On modern architectures with hierarchical memory, the cost of loading and storing input matrix elements tends to dominate the cost of arithmetic. On a single machine this is the amount of data transferred between RAM and cache, while on a distributed memory multi-node machine it is the amount transferred between nodes; in either case it is called the *communication bandwidth*. The naïve algorithm using three nested loops uses $\Omega(n^3)$ communication bandwidth.

Cannon's algorithm, also known as the *2D algorithm*, partitions each input matrix into a block matrix whose elements are submatrices of size $\sqrt{M/3}$ by $\sqrt{M/3}$, where M is the size of fast memory.^[10] The naïve algorithm is then used over the block matrices, computing products of submatrices entirely in fast memory. This reduces communication bandwidth to $O(n^3/\sqrt{M})$, which is asymptotically optimal (for algorithms performing $\Omega(n^3)$ computation).

In a distributed setting with p processors arranged in a \sqrt{p} by \sqrt{p} 2D mesh, one submatrix of the result can be assigned to each processor, and the product can be computed with each processor transmitting $O(n^2/\sqrt{p})$ words, which is asymptotically optimal assuming that each node stores the minimum $O(n^2/p)$ elements. This can be improved by the *3D algorithm*, which arranges the processors in a 3D cube mesh, assigning every product of two input submatrices to a single processor. The result submatrices are then generated by performing a reduction over each row. This algorithm transmits $O(n^2/p^{2/3})$ words per processor, which is asymptotically optimal. However, this requires replicating each input matrix element $p^{1/3}$ times, and so requires a factor of $p^{1/3}$ more memory than is needed to store the inputs. This algorithm can be combined with Strassen to further reduce runtime. "2.5D" algorithms provide a continuous tradeoff between memory usage and communication bandwidth.



Other forms of multiplication

Some other ways to multiply two matrices are given below, in fact simpler than the definition above.

Hadamard product

For two matrices of the same dimensions, there is the **Hadamard product**, also known as the **element-wise product**, **pointwise product**, **entrywise product** and the **Schur product**. For two matrices \mathbf{A} and \mathbf{B} of the same dimensions, the Hadamard product $\mathbf{A} \circ \mathbf{B}$ is a matrix of the same dimensions, the i, j element of \mathbf{A} is multiplied with the i, j element of \mathbf{B} , that is:

$$(\mathbf{A} \circ \mathbf{B})_{ij} = A_{ij}B_{ij},$$

displayed fully:

$$\mathbf{A} \circ \mathbf{B} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix} \circ \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nm} \end{pmatrix} = \begin{pmatrix} A_{11}B_{11} & A_{12}B_{12} & \cdots & A_{1m}B_{1m} \\ A_{21}B_{21} & A_{22}B_{22} & \cdots & A_{2m}B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1}B_{n1} & A_{n2}B_{n2} & \cdots & A_{nm}B_{nm} \end{pmatrix}$$

This operation is identical to many multiplying ordinary numbers (mn of them) all at once; thus the Hadamard product is commutative, associative and distributive over entrywise addition. It is also a principal submatrix of the Kronecker product. It appears in lossy compression algorithms such as JPEG.

Frobenius product

The **Frobenius inner product**, sometimes denoted $\mathbf{A} : \mathbf{B}$, is the component-wise inner product of two matrices as though they are vectors. It is also the sum of the entries of the Hadamard product. Explicitly,

$$\mathbf{A} : \mathbf{B} = \sum_{i,j} A_{ij}B_{ij} = \text{vec}(\mathbf{A})^\top \text{vec}(\mathbf{B}) = \text{tr}(\mathbf{A}^\top \mathbf{B}) = \text{tr}(\mathbf{A}\mathbf{B}^\top),$$

where "tr" denotes the trace of a matrix and vec denotes vectorization. This inner product induces the Frobenius norm.

Kronecker product

For two matrices \mathbf{A} and \mathbf{B} of any different dimensions $m \times n$ and $p \times q$ respectively (no constraints on the dimensions of each matrix), the **Kronecker product** denoted $\mathbf{A} \otimes \mathbf{B}$ is a matrix with dimensions $mp \times nq$, which has elements^[citation needed]

$$(\mathbf{A} \otimes \mathbf{B})_{ij} = A_{1+\lfloor \frac{i-1}{p} \rfloor, 1+\lfloor \frac{j-1}{q} \rfloor} B_{1+(i-1) \bmod p, 1+(j-1) \bmod q},$$

where $\lfloor \cdot \rfloor$ represents the floor function.

Explicitly:

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} A_{11}\mathbf{B} & A_{12}\mathbf{B} & \cdots & A_{1n}\mathbf{B} \\ A_{21}\mathbf{B} & A_{22}\mathbf{B} & \cdots & A_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1}\mathbf{B} & A_{m2}\mathbf{B} & \cdots & A_{mn}\mathbf{B} \end{pmatrix}.$$

This is the application of the more general tensor product applied to matrices.

Notes

- [1] Encyclopaedia of Physics (2nd Edition), R.G. Lerner, G.L. Trigg, VHC publishers, 1991, ISBN (Verlagsgesellschaft) 3-527-26954-1, ISBN (VHC Inc.) 0-89573-752-3
- [2] McGraw Hill Encyclopaedia of Physics (2nd Edition), C.B. Parker, 1994, ISBN 0-07-051400-3
- [3] *Linear Algebra* (4th Edition), S. Lipschutz, M. Lipson, Schaum's Outlines, McGraw Hill (USA), 2009, ISBN 978-0-07-154352-1
- [4] *Mathematical methods for physics and engineering*, K.F. Riley, M.P. Hobson, S.J. Bence, Cambridge University Press, 2010, ISBN 978-0-521-86153-3
- [5] *Mathematical methods for physics and engineering*, K.F. Riley, M.P. Hobson, S.J. Bence, Cambridge University Press, 2010, ISBN 978-0-521-86153-3
- [6] Press 2007, p. 108.
- [7] The original algorithm was presented by Don Coppersmith and Shmuel Winograd in 1990, has an asymptotic complexity of .
- [8] Robinson, 2005.
- [9] Alon, Shpilka, Umans, On Sunflowers and Matrix Multiplication (<http://eccc.hpi-web.de/report/2011/067/>)
- [10] Lynn Elliot Cannon, *A cellular computer to implement the Kalman Filter Algorithm* (<http://portal.acm.org/citation.cfm?coll=GUIDE&dl=GUIDE&id=905686>), Technical report, Ph.D. Thesis, Montana State University, 14 July 1969.

References

- Henry Cohn, Robert Kleinberg, Balazs Szegedy, and Chris Umans. Group-theoretic Algorithms for Matrix Multiplication. arXiv:math.GR/0511460. *Proceedings of the 46th Annual Symposium on Foundations of Computer Science*, 23–25 October 2005, Pittsburgh, PA, IEEE Computer Society, pp. 379–388.
- Henry Cohn, Chris Umans. A Group-theoretic Approach to Fast Matrix Multiplication. arXiv:math.GR/0307321. *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 11–14 October 2003, Cambridge, MA, IEEE Computer Society, pp. 438–449.
- Coppersmith, D., Winograd S., *Matrix multiplication via arithmetic progressions*, J. Symbolic Comput. 9, p. 251-280, 1990.
- Horn, Roger A.; Johnson, Charles R. (1985), *Matrix Analysis*, Cambridge University Press, ISBN 978-0-521-38632-6
- Horn, Roger A.; Johnson, Charles R. (1991), *Topics in Matrix Analysis*, Cambridge University Press, ISBN 978-0-521-46713-1
- Knuth, D.E., *The Art of Computer Programming Volume 2: Seminumerical Algorithms*. Addison-Wesley Professional; 3 edition (November 14, 1997). ISBN 978-0-201-89684-8. pp. 501.
- Press, William H.; Flannery, Brian P.; Teukolsky, Saul A.; Vetterling, William T. (2007), *Numerical Recipes: The Art of Scientific Computing* (3rd ed.), Cambridge University Press, ISBN 978-0-521-88068-8.
- Ran Raz. On the complexity of matrix product. In Proceedings of the thirty-fourth annual ACM symposium on Theory of computing. ACM Press, 2002. doi: 10.1145/509907.509932 (<http://dx.doi.org/10.1145/509907.509932>).
- Robinson, Sara, *Toward an Optimal Algorithm for Matrix Multiplication*, SIAM News 38(9), November 2005. PDF (<http://www.siam.org/pdf/news/174.pdf>)
- Strassen, Volker, *Gaussian Elimination is not Optimal*, Numer. Math. 13, p. 354-356, 1969.
- Styan, George P. H. (1973), "Hadamard Products and Multivariate Statistical Analysis", *Linear Algebra and its Applications* 6: 217–240, doi: 10.1016/0024-3795(73)90023-2 ([http://dx.doi.org/10.1016/0024-3795\(73\)90023-2](http://dx.doi.org/10.1016/0024-3795(73)90023-2))
- Vassilevska Williams, Virginia, *Multiplying matrices faster than Coppersmith-Winograd*, Manuscript, May 2012. PDF (<http://www.cs.stanford.edu/~virgi/matrixmult-f.pdf>)

External links

- How to Multiply Matrices (<http://www.mathsisfun.com/algebra/matrix-multiplying.html>)
- The Simultaneous Triple Product Property and Group-theoretic Results for the Exponent of Matrix Multiplication
- WIMS Online Matrix Multiplier (http://wims.unice.fr/~wims/en_tool~linear~matmult.html)
- Matrix Multiplication Problems (<http://ceee.rice.edu/Books/LA/mult/mult4.html#TOP>)
- Block Matrix Multiplication Problems (<http://www.gordon-taft.net/MatrixMultiplication.html>)
- Wijesuriya, Viraj B., *Daniweb: Sample Code for Matrix Multiplication using MPI Parallel Programming Approach* (<http://www.daniweb.com/forums/post1428830.html#post1428830>), retrieved 2010-12-29
- Linear algebra: matrix operations (http://www.umat.feec.vutbr.cz/~novakm/algebra_matic/en) Multiply or add matrices of a type and with coefficients you choose and see how the result was computed.
- Visual Matrix Multiplication (http://www.wefoundland.com/project/Visual_Matrix_Multiplication) An interactive app for learning matrix multiplication.
- Matrix Multiplication in Java – Dr. P. Viry (http://www.ateji.com/px/whitepapers/Ateji_PX_MatMult_Whitepaper_v1.2.pdf?phpMyAdmin=95wsvAC1wsqrAq3j,M3duZU3UJ7)

Transpose

This article is about the transpose of a matrix. For other uses, see Transposition

Note that this article assumes that matrices are taken over a commutative ring. These results may not hold in the non-commutative case.

In linear algebra, the **transpose** of a matrix **A** is another matrix **A^T** (also written **A'**, **A^{tr}**, **A** or **A[†]**) created by any one of the following equivalent actions:

- reflect **A** over its main diagonal (which runs from top-left to bottom-right) to obtain **A^T**
- write the rows of **A** as the columns of **A^T**
- write the columns of **A** as the rows of **A^T**

Formally, the *i* th row, *j* th column element of **A^T** is the *j* th row, *i* th column element of **A**:

$$[\mathbf{A}^T]_{ij} = [\mathbf{A}]_{ji}$$

If **A** is an *m* × *n* matrix then **A^T** is an *n* × *m* matrix.

The transpose of a matrix was introduced in 1858 by the British mathematician Arthur Cayley.^[1]

A

$\left[\begin{array}{cc} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{array} \right]$

The transpose **A^T** of a matrix **A** can be obtained by reflecting the elements along its main diagonal. Repeating the process on the transposed matrix returns the elements to their original position.

Examples

- $\begin{bmatrix} 1 & 2 \end{bmatrix}^T = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$
- $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$
- $\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$

Properties

For matrices **A**, **B** and scalar *c* we have the following properties of transpose:

1. $(\mathbf{A}^T)^T = \mathbf{A}$

The operation of taking the transpose is an involution (self-inverse).

- $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$

The transpose respects addition.

- $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$

Note that the order of the factors reverses. From this one can deduce that a square matrix **A** is invertible if and only if **A^T** is invertible, and in this case we have $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$. By induction this result extends to the general case of multiple matrices, where we find that $(\mathbf{A}_1 \mathbf{A}_2 \dots \mathbf{A}_{k-1} \mathbf{A}_k)^T = \mathbf{A}_k^T \mathbf{A}_{k-1}^T \dots \mathbf{A}_2^T \mathbf{A}_1^T$.

- $(c\mathbf{A})^T = c\mathbf{A}^T$

The transpose of a scalar is the same scalar. Together with (2), this states that the transpose is a linear map from the space of *m* × *n* matrices to the space of all *n* × *m* matrices.

- $\det(\mathbf{A}^T) = \det(\mathbf{A})$

The determinant of a square matrix is the same as that of its transpose.

- The dot product of two column vectors \mathbf{a} and \mathbf{b} can be computed as

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b},$$

which is written as $\mathbf{a}_i \mathbf{b}^i$ in Einstein notation.

2. If \mathbf{A} has only real entries, then $\mathbf{A}^T \mathbf{A}$ is a positive-semidefinite matrix.

3. $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$

The transpose of an invertible matrix is also invertible, and its inverse is the transpose of the inverse of the original matrix. The notation \mathbf{A}^{-T} is sometimes used to represent either of these equivalent expressions.

- If \mathbf{A} is a square matrix, then its eigenvalues are equal to the eigenvalues of its transpose since they share the same Characteristic polynomial.

Special transpose matrices

A square matrix whose transpose is equal to itself is called a symmetric matrix; that is, \mathbf{A} is symmetric if

$$\mathbf{A}^T = \mathbf{A}.$$

A square matrix whose transpose is equal to its negative is called a skew-symmetric matrix; that is, \mathbf{A} is skew-symmetric if

$$\mathbf{A}^T = -\mathbf{A}.$$

A square complex matrix whose transpose is equal to the matrix with every entry replaced by its complex conjugate is called a Hermitian matrix (equivalent to the matrix being equal to its conjugate transpose); that is, \mathbf{A} is Hermitian if

$$\mathbf{A}^T = \mathbf{A}^*.$$

A square complex matrix whose transpose is equal to the negation of its complex conjugate is called a skew-Hermitian matrix; that is, \mathbf{A} is skew-Hermitian if

$$\mathbf{A}^T = -\mathbf{A}^*.$$

A square matrix whose transpose is equal to its inverse is called an orthogonal matrix; that is, \mathbf{A} is orthogonal if

$$\mathbf{A}^T = \mathbf{A}^{-1}.$$

Transpose of a linear map

The transpose may be defined using a coordinate-free approach:

If $f: V \rightarrow W$ is a linear map between vector spaces V and W with respective dual spaces V^* and W^* , the *transpose* of f is the linear map ${}^t f: W^* \rightarrow V^*$ that satisfies

$${}^t f(\phi) = \phi \circ f \quad \forall \phi \in W^*.$$

The definition of the transpose may be seen to be independent of any bilinear form on the vector spaces, unlike the adjoint (below).

If the matrix A describes a linear map with respect to bases of V and W , then the matrix A^T describes the transpose of that linear map with respect to the dual bases.

Transpose of a bilinear form

Every linear map to the dual space $f: V \rightarrow V^*$ defines a bilinear form $B: V \times V \rightarrow F$, with the relation $B(\mathbf{v}, \mathbf{w}) = f(\mathbf{v})(\mathbf{w})$. By defining the transpose of this bilinear form as the bilinear form tB defined by the transpose ${}^t f: V^{**} \rightarrow V^*$ i.e. ${}^tB(\mathbf{w}, \mathbf{v}) = {}^t f(\mathbf{w})(\mathbf{v})$, we find that $B(\mathbf{v}, \mathbf{w}) = {}^tB(\mathbf{w}, \mathbf{v})$.

Adjoint

If the vector spaces V and W have respective nondegenerate bilinear forms B_V and B_W , a concept closely related to the transpose – the *adjoint* – may be defined:

If $f: V \rightarrow W$ is a linear map between vector spaces V and W , we define g as the *adjoint* of f if $g: W \rightarrow V$ satisfies

$$B_V(v, g(w)) = B_W(f(v), w) \quad \forall v \in V, w \in W.$$

These bilinear forms define an isomorphism between V and V^* , and between W and W^* , resulting in an isomorphism between the transpose and adjoint of f . The matrix of the adjoint of a map is the transposed matrix only if the bases are orthonormal with respect to their bilinear forms. In this context, many authors use the term transpose to refer to the adjoint as defined here.

The adjoint allows us to consider whether $g: W \rightarrow V$ is equal to $f^{-1}: W \rightarrow V$. In particular, this allows the orthogonal group over a vector space V with a quadratic form to be defined without reference to matrices (nor the components thereof) as the set of all linear maps $V \rightarrow V$ for which the adjoint equals the inverse.

Over a complex vector space, one often works with sesquilinear forms (conjugate-linear in one argument) instead of bilinear forms. The Hermitian adjoint of a map between such spaces is defined similarly, and the matrix of the Hermitian adjoint is given by the conjugate transpose matrix if the bases are orthonormal.

Implementation of matrix transposition on computers

On a computer, one can often avoid explicitly transposing a matrix in memory by simply accessing the same data in a different order. For example, software libraries for linear algebra, such as BLAS, typically provide options to specify that certain matrices are to be interpreted in transposed order to avoid the necessity of data movement.

However, there remain a number of circumstances in which it is necessary or desirable to physically reorder a matrix in memory to its transposed ordering. For example, with a matrix stored in row-major order, the rows of the matrix are contiguous in memory and the columns are discontinuous. If repeated operations need to be performed on the columns, for example in a fast Fourier transform algorithm, transposing the matrix in memory (to make the columns contiguous) may improve performance by increasing memory locality.

Ideally, one might hope to transpose a matrix with minimal additional storage. This leads to the problem of transposing an $n \times m$ matrix in-place, with $O(1)$ additional storage or at most storage much less than mn . For $n \neq m$, this involves a complicated permutation of the data elements that is non-trivial to implement in-place. Therefore efficient in-place matrix transposition has been the subject of numerous research publications in computer science, starting in the late 1950s, and several algorithms have been developed.

References

- [1] Arthur Cayley (1858) "A memoir on the theory of matrices," (<http://books.google.com/books?id=f1FFAAAACAAJ&pg=PA31#v=onepage&q&f=false>) *Philosophical Transactions of the Royal Society of London*, **148** : 17-37. The transpose (or "transposition") is defined on page 31.

External links

- MIT Linear Algebra Lecture on Matrix Transposes (<http://ocw.mit.edu/OcwWeb/Mathematics/18-06Spring-2005/VideoLectures/detail/lecture05.htm>)
- Transpose (<http://mathworld.wolfram.com/Transpose.html>), mathworld.wolfram.com
- Transpose (<http://planetmath.org/encyclopedia/Transpose.html>), planetmath.org
- Khan Academy introduction to matrix transposes (http://khanexercises.appspot.com/video?v=2t0003_sxtU)

Determinant

In linear algebra, the **determinant** is a value associated with a square matrix. It can be computed from the entries of the matrix by a specific arithmetic expression, while other ways to determine its value exist as well. The determinant provides important information about a matrix of coefficients of a system of linear equations, or about a matrix that corresponds to a linear transformation of a vector space. In the first case the system has a unique solution exactly when the determinant is nonzero; when the determinant is zero there are either no solutions or many solutions. In the second case the transformation has an inverse operation exactly when the determinant is nonzero. A geometric interpretation can be given to the value of the determinant of a square matrix with real entries: the absolute value of the determinant gives the scale factor by which area or volume (or a higher dimensional analogue) is multiplied under the associated linear transformation, while its sign indicates whether the transformation preserves orientation. Thus a 2×2 matrix with determinant -2 , when applied to a region of the plane with finite area, will transform that region into one with twice the area, while reversing its orientation.

Determinants occur throughout mathematics. The use of determinants in calculus includes the Jacobian determinant in the substitution rule for integrals of functions of several variables. They are used to define the characteristic polynomial of a matrix that is an essential tool in eigenvalue problems in linear algebra. In some cases they are used just as a compact notation for expressions that would otherwise be unwieldy to write down.

The determinant of a matrix \mathbf{A} is denoted $\det(\mathbf{A})$, $\det \mathbf{A}$, or $|\mathbf{A}|$. In the case where the matrix entries are written out in full, the determinant is denoted by surrounding the matrix entries by vertical bars instead of the brackets or parentheses of the matrix. For instance, the determinant of the matrix

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

is written

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}$$

and has the value

$$(aei + bfg + cdh) - (ceg + bdi + afh).$$

Although most often used for matrices whose entries are real or complex numbers, the definition of the determinant only involves addition, subtraction and multiplication, and so it can be defined for square matrices with entries taken from any commutative ring. Thus for instance the determinant of a matrix with integer coefficients will be an

integer, and the matrix has an inverse with integer coefficients if and only if this determinant is 1 or -1 (these being the only invertible elements of the integers). For square matrices with entries in a non-commutative ring, for instance the quaternions, there is no unique definition for the determinant, and no definition that has all the usual properties of determinants over commutative rings.

Definition

There are various ways to define the determinant of a square matrix A , i.e. one with the same number of rows and columns. Perhaps the most natural way is expressed in terms of the columns of the matrix. If we write an $n \times n$ matrix in terms of its column vectors

$$A = [a_1, a_2, \dots, a_n]$$

where the a_j are vectors of size n , then the determinant of A is defined so that

$$\det [a_1, \dots, ba_j + cv, \dots, a_n] = b \det(A) + c \det [a_1, \dots, v, \dots, a_n]$$

$$\det [a_1, \dots, a_j, a_{j+1}, \dots, a_n] = - \det [a_1, \dots, a_{j+1}, a_j, \dots, a_n]$$

$$\det(I) = 1$$

where b and c are scalars, v is any vector of size n and I is the identity matrix of size n . These equations say that the determinant is a linear function of each column, that interchanging adjacent columns reverses the sign of the determinant, and that the determinant of the identity matrix is 1. These properties mean that the determinant is an alternating multilinear function of the columns that maps the identity matrix to the underlying unit scalar. These suffice to uniquely calculate the determinant of any square matrix. Provided the underlying scalars form a field (more generally, a commutative ring with unity), the definition below shows that such a function exists, and it can be shown to be unique.^[1]

Equivalently, the determinant can be expressed as a sum of products of entries of the matrix where each product has n terms and the coefficient of each product is -1 or 1 or 0 according to a given rule: it is a polynomial expression of the matrix entries. This expression grows rapidly with the size of the matrix (an $n \times n$ matrix contributes $n!$ terms), so it will first be given explicitly for the case of 2×2 matrices and 3×3 matrices, followed by the rule for arbitrary size matrices, which subsumes these two cases.

Assume A is a square matrix with n rows and n columns, so that it can be written as

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix}.$$

The entries can be numbers or expressions (as happens when the determinant is used to define a characteristic polynomial); the definition of the determinant depends only on the fact that they can be added and multiplied together in a commutative manner.

The determinant of A is denoted as $\det(A)$, or it can be denoted directly in terms of the matrix entries by writing enclosing bars instead of brackets:

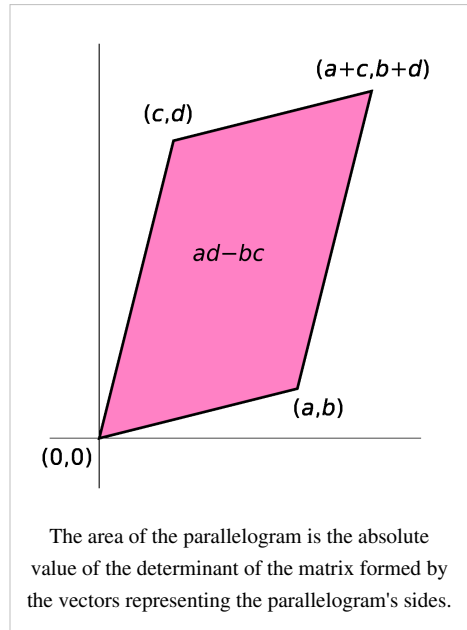
$$\begin{vmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{vmatrix}.$$

2 × 2 matrices

The determinant of a 2 × 2 matrix is defined by

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

If the matrix entries are real numbers, the matrix A can be used to represent two linear mappings: one that maps the standard basis vectors to the rows of A , and one that maps them to the columns of A . In either case, the images of the basis vectors form a parallelogram that represents the image of the unit square under the mapping. The parallelogram defined by the rows of the above matrix is the one with vertices at $(0, 0)$, (a, b) , $(a + c, b + d)$, and (c, d) , as shown in the accompanying diagram. The absolute value of $ad - bc$ is the area of the parallelogram, and thus represents the scale factor by which areas are transformed by A . (The parallelogram formed by the columns of A is in general a different parallelogram, but since the determinant is symmetric with respect to rows and columns, the area will be the same.)



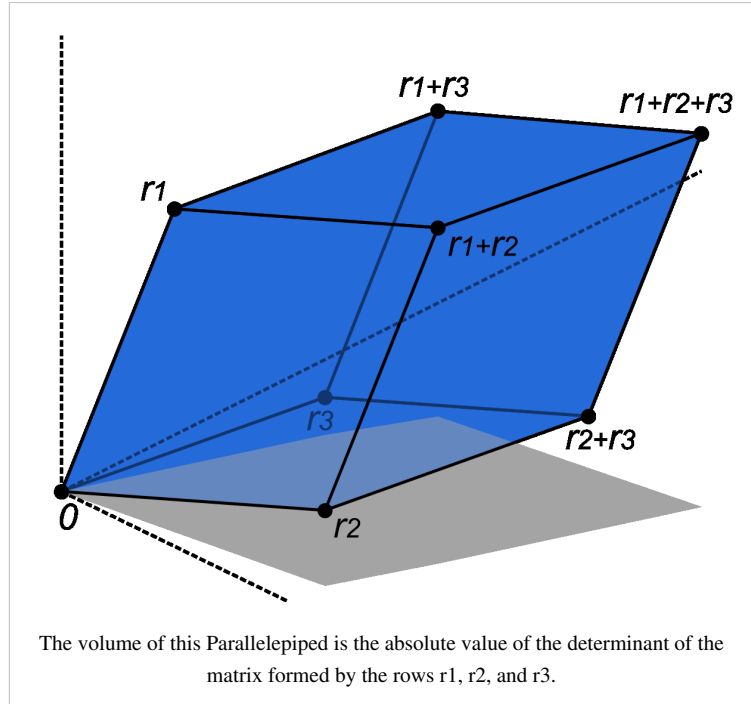
The absolute value of the determinant together with the sign becomes the *oriented area* of the parallelogram. The oriented area is the same as the usual area, except that it is negative when the angle from the first to the second vector defining the parallelogram turns in a clockwise direction (which is opposite to the direction one would get for the identity matrix).

Thus the determinant gives the scaling factor and the orientation induced by the mapping represented by A . When the determinant is equal to one, the linear mapping defined by the matrix is equi-areal and orientation-preserving.

The object known as the *bivector* is related to these ideas. In 2d, it can be interpreted as an *oriented plane segment* formed by imagining two vectors each with origin $(0, 0)$, and coordinates (a, b) and (c, d) . The bivector magnitude (denoted $(a, b) \wedge (c, d)$) is the *signed area*, which is also the determinant $ad - bc$.^[2]

3 × 3 matrices

The determinant of a 3×3 matrix is defined by



$$\begin{aligned} \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} &= a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= a(ei - fh) - b(di - fg) + c(dh - eg) \\ &= aei + bfg + cdh - ceg - bdi - afh. \end{aligned}$$

The rule of Sarrus is a mnemonic for the 3x3 matrix determinant: the sum of the products of three diagonal north-west to south-east lines of matrix elements, minus the sum of the products of three diagonal south-west to north-east lines of elements, when the copies of the first two columns of the matrix are written beside it as in the illustration. This scheme for calculating the determinant of a 3 × 3 matrix does not carry over into higher dimensions.

n × n matrices

The determinant of a matrix of arbitrary size can be defined by the Leibniz formula or the Laplace formula.

The Leibniz formula for the determinant of an n × n matrix **A** is

$$\det(A) = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^n A_{i,\sigma_i}.$$

Here the sum is computed over all permutations σ of the set $\{1, 2, \dots, n\}$. A permutation is a function that reorders this set of integers. The value in the i th position after the reordering σ is denoted σ_i . For example, for $n = 3$, the original sequence 1, 2, 3 might be reordered to $\sigma = [2, 3, 1]$, with $\sigma_1 = 2$, $\sigma_2 = 3$, and $\sigma_3 = 1$. The set of all such permutations (also known as the symmetric group on n elements) is denoted S_n . For each permutation σ , $\text{sgn}(\sigma)$ denotes the signature of σ , a value that is +1 whenever the reordering given by σ can be achieved by successively interchanging two entries an even number of times, and -1 whenever it can be achieved by an odd number of such interchanges.

In any of the $n!$ summands, the term

$$\prod_{i=1}^n A_{i,\sigma_i}$$

is notation for the product of the entries at positions (i, σ_i) , where i ranges from 1 to n :

$$A_{1,\sigma_1} \cdot A_{2,\sigma_2} \cdots A_{n,\sigma_n}.$$

For example, the determinant of a 3×3 matrix A ($n = 3$) is

$$\begin{aligned} \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^n A_{i,\sigma_i} &= \text{sgn}([1, 2, 3]) \prod_{i=1}^n A_{i,[1,2,3]_i} + \text{sgn}([1, 3, 2]) \prod_{i=1}^n A_{i,[1,3,2]_i} + \text{sgn}([2, 1, 3]) \prod_{i=1}^n A_{i,[2,1,3]_i} \\ &+ \text{sgn}([2, 3, 1]) \prod_{i=1}^n A_{i,[2,3,1]_i} + \text{sgn}([3, 1, 2]) \prod_{i=1}^n A_{i,[3,1,2]_i} + \text{sgn}([3, 2, 1]) \prod_{i=1}^n A_{i,[3,2,1]_i} \\ &= \prod_{i=1}^n A_{i,[1,2,3]_i} - \prod_{i=1}^n A_{i,[1,3,2]_i} - \prod_{i=1}^n A_{i,[2,1,3]_i} + \prod_{i=1}^n A_{i,[2,3,1]_i} + \prod_{i=1}^n A_{i,[3,1,2]_i} - \prod_{i=1}^n A_{i,[3,2,1]_i} \\ &= A_{1,1}A_{2,2}A_{3,3} - A_{1,1}A_{2,3}A_{3,2} - A_{1,2}A_{2,1}A_{3,3} + A_{1,2}A_{2,3}A_{3,1} \\ &\quad + A_{1,3}A_{2,1}A_{3,2} - A_{1,3}A_{2,2}A_{3,1}. \end{aligned}$$

Levi-Civita symbol

It is sometimes useful to extend the Leibniz formula to a summation in which not only permutations, but all sequences of n indices in the range $1, \dots, n$ occur, ensuring that the contribution of a sequence will be zero unless it denotes a permutation. Thus the totally antisymmetric Levi-Civita symbol $\varepsilon_{i_1, \dots, i_n}$ extends the signature of a permutation, by setting $\varepsilon_{\sigma(1), \dots, \sigma(n)} = \text{sgn}(\sigma)$ for any permutation σ of n , and $\varepsilon_{i_1, \dots, i_n} = 0$ when no permutation σ exists such that $\sigma(j) = i_j$ for $j = 1, \dots, n$ (or equivalently, whenever some pair of indices are equal). The determinant for an $n \times n$ matrix can then be expressed using an n -fold summation as

$$\det A = \sum_{i_1, i_2, \dots, i_n=1}^n \varepsilon_{i_1 \dots i_n} a_{1,i_1} \cdots a_{n,i_n}.$$

Properties of the determinant

The determinant has many properties. Some basic properties of determinants are:

1. $\det(I_n) = 1$ where I_n is the $n \times n$ identity matrix.
2. $\det(A^T) = \det(A)$.
3. $\det(A^{-1}) = \frac{1}{\det(A)} = \det(A)^{-1}$.
4. For square matrices A and B of equal size,

$$\det(AB) = \det(A) \det(B).$$

- $\det(cA) = c^n \det(A)$ for an $n \times n$ matrix.
- 2. If A is a triangular matrix, i.e. $a_{i,j} = 0$ whenever $i > j$ or, alternatively, whenever $i < j$, then its determinant equals the product of the diagonal entries:

$$\det(A) = a_{1,1}a_{2,2} \cdots a_{n,n} = \prod_{i=1}^n a_{i,i}.$$

This can be deduced from some of the properties below, but it follows most easily directly from the Leibniz formula (or from the Laplace expansion), in which the identity permutation is the only one that gives a non-zero contribution.

A number of additional properties relate to the effects on the determinant of changing particular rows or columns:

- Viewing an $n \times n$ matrix as being composed of n columns, the determinant is an n -linear function. This means that if one column of a matrix A is written as a sum $v + w$ of two column vectors, and all other columns are left unchanged, then the determinant of A is the sum of the determinants of the matrices obtained from A by replacing the column by v and then by w (and a similar relation holds when writing a column as a scalar multiple of a column vector).

2. This n -linear function is an alternating form. This means that whenever two columns of a matrix are identical, or more generally some column can be expressed as a linear combination of the other columns (i.e. the columns of the matrix form a linearly dependent set), its determinant is 0.

Properties 1, 7 and 8 — which all follow from the Leibniz formula — completely characterize the determinant; in other words the determinant is the unique function from $n \times n$ matrices to scalars that is n -linear alternating in the columns, and takes the value 1 for the identity matrix (this characterization holds even if scalars are taken in any given commutative ring). To see this it suffices to expand the determinant by multi-linearity in the columns into a (huge) linear combination of determinants of matrices in which each column is a standard basis vector. These determinants are either 0 (by property 8) or else ± 1 (by properties 1 and 11 below), so the linear combination gives the expression above in terms of the Levi-Civita symbol. While less technical in appearance, this characterization cannot entirely replace the Leibniz formula in defining the determinant, since without it the existence of an appropriate function is not clear. For matrices over non-commutative rings, properties 7 and 8 are incompatible for $n \geq 2$,^[3] so there is no good definition of the determinant in this setting.

Property 2 above implies that properties for columns have their counterparts in terms of rows:

- Viewing an $n \times n$ matrix as being composed of n rows, the determinant is an n -linear function.
2. This n -linear function is an alternating form: whenever two rows of a matrix are identical, its determinant is 0.
 3. Interchanging two columns of a matrix multiplies its determinant by -1 . This follows from properties 7 and 8 (it is a general property of multilinear alternating maps). Iterating gives that more generally a permutation of the columns multiplies the determinant by the sign of the permutation. Similarly a permutation of the rows multiplies the determinant by the sign of the permutation.
 4. Adding a scalar multiple of one column to *another* column does not change the value of the determinant. This is a consequence of properties 7 and 8: by property 7 the determinant changes by a multiple of the determinant of a matrix with two equal columns, which determinant is 0 by property 8. Similarly, adding a scalar multiple of one row to another row leaves the determinant unchanged.

These properties can be used to facilitate the computation of determinants by simplifying the matrix to the point where the determinant can be determined immediately. Specifically, for matrices with coefficients in a field, properties 11 and 12 can be used to transform any matrix into a triangular matrix, whose determinant is given by property 6; this is essentially the method of Gaussian elimination.

For example, the determinant of

$$A = \begin{bmatrix} -2 & 2 & -3 \\ -1 & 1 & 3 \\ 2 & 0 & -1 \end{bmatrix}$$

can be computed using the following matrices:

$$B = \begin{bmatrix} -2 & 2 & -3 \\ 0 & 0 & 4.5 \\ 2 & 0 & -1 \end{bmatrix}, \quad C = \begin{bmatrix} -2 & 2 & -3 \\ 0 & 0 & 4.5 \\ 0 & 2 & -4 \end{bmatrix}, \quad D = \begin{bmatrix} -2 & 2 & -3 \\ 0 & 2 & -4 \\ 0 & 0 & 4.5 \end{bmatrix}.$$

Here, B is obtained from A by adding $-1/2 \times$ the first row to the second, so that $\det(A) = \det(B)$. C is obtained from B by adding the first to the third row, so that $\det(C) = \det(B)$. Finally, D is obtained from C by exchanging the second and third row, so that $\det(D) = -\det(C)$. The determinant of the (upper) triangular matrix D is the product of its entries on the main diagonal: $(-2) \cdot 2 \cdot 4.5 = -18$. Therefore $\det(A) = -\det(D) = +18$.

Multiplicativity and matrix groups

The determinant of a matrix product of square matrices equals the product of their determinants:

$$\det(AB) = \det(A) \det(B).$$

Thus the determinant is a *multiplicative map*. This property is a consequence of the characterization given above of the determinant as the unique n -linear alternating function of the columns with value 1 on the identity matrix, since the function $M_n(K) \rightarrow K$ that maps $M \mapsto \det(AM)$ can easily be seen to be n -linear and alternating in the columns of M , and takes the value $\det(A)$ at the identity. The formula can be generalized to (square) products of rectangular matrices, giving the Cauchy–Binet formula, which also provides an independent proof of the multiplicative property.

The determinant $\det(A)$ of a matrix A is non-zero if and only if A is invertible or, yet another equivalent statement, if its rank equals the size of the matrix. If so, the determinant of the inverse matrix is given by

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

In particular, products and inverses of matrices with determinant one still have this property. Thus, the set of such matrices (of fixed size n) form a group known as the special linear group. More generally, the word "special" indicates the subgroup of another matrix group of matrices of determinant one. Examples include the special orthogonal group (which if n is 2 or 3 consists of all rotation matrices), and the special unitary group.

Laplace's formula and the adjugate matrix

Laplace's formula expresses the determinant of a matrix in terms of its minors. The minor $M_{i,j}$ is defined to be the determinant of the $(n-1) \times (n-1)$ -matrix that results from A by removing the i th row and the j th column. The expression $(-1)^{i+j}M_{i,j}$ is known as cofactor. The determinant of A is given by

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} a_{i,j} M_{i,j} = \sum_{i=1}^n (-1)^{i+j} a_{i,j} M_{i,j}.$$

Calculating $\det(A)$ by means of that formula is referred to as expanding the determinant along a row or column. For the example 3×3 matrix

$$A = \begin{bmatrix} -2 & 2 & -3 \\ -1 & 1 & 3 \\ 2 & 0 & -1 \end{bmatrix},$$

Laplace expansion along the second column ($j = 2$, the sum runs over i) yields:

$$\begin{aligned} \det(A) &= (-1)^{1+2} \cdot 2 \cdot \det \begin{bmatrix} -1 & 3 \\ 2 & -1 \end{bmatrix} + (-1)^{2+2} \cdot 1 \cdot \det \begin{bmatrix} -2 & -3 \\ 2 & -1 \end{bmatrix} + (-1)^{3+2} \cdot 0 \cdot \det \begin{bmatrix} -2 & -3 \\ -1 & 3 \end{bmatrix} \\ &= (-2) \cdot ((-1) \cdot (-1) - 2 \cdot 3) + 1 \cdot ((-2) \cdot (-1) - 2 \cdot (-3)) \\ &= (-2) \cdot (-5) + 8 = 18. \end{aligned}$$

However, Laplace expansion is efficient for small matrices only.

The adjugate matrix $\text{adj}(A)$ is the transpose of the matrix consisting of the cofactors, i.e.,

$$(\text{adj}(A))_{i,j} = (-1)^{i+j} M_{j,i}.$$

Sylvester's determinant theorem

Sylvester's determinant theorem states that for A , an $m \times n$ matrix, and B , an $n \times m$ matrix (so that A and B have dimensions allowing them to be multiplied in either order):

$$\det(I_m + AB) = \det(I_n + BA),$$

where I_m and I_n are the $m \times m$ and $n \times n$ identity matrices, respectively.

From this general result several consequences follow.

(a) For the case of column vector c and row vector r , each with m components, the formula allows quick calculation of the determinant of a matrix that differs from the identity matrix by a matrix of rank 1:

$$\det(I_m + cr) = 1 + rc.$$

(b) More generally,^[4] for any invertible $m \times m$ matrix X ,

$$\det(X + AB) = \det(X) \det(I_n + BX^{-1}A),$$

(c) For a column and row vector as above, $\det(X + cr) = \det(X)(1 + rX^{-1}c)$.

Properties of the determinant in relation to other notions

Relation to eigenvalues and trace

Determinants can be used to find the eigenvalues of the matrix A : they are the solutions of the characteristic equation

$$\det(A - xI) = 0,$$

where I is the identity matrix of the same dimension as A . Conversely, $\det(A)$ is the product of the eigenvalues of A , counted with their algebraic multiplicities. The product of all non-zero eigenvalues is referred to as pseudo-determinant.

An Hermitian matrix is positive definite if all its eigenvalues are positive. Sylvester's criterion asserts that this is equivalent to the determinants of the submatrices

$$A_k := \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ a_{2,1} & a_{2,2} & \dots & a_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} & a_{k,2} & \dots & a_{k,k} \end{bmatrix}$$

being positive, for all k between 1 and n .

The trace $\text{tr}(A)$ is by definition the sum of the diagonal entries of A and also equals the sum of the eigenvalues. Thus, for complex matrices A ,

$$\det(\exp(A)) = \exp(\text{tr}(A))$$

or, for real matrices A ,

$$\text{tr}(A) = \log(\det(\exp(A))).$$

Here $\exp(A)$ denotes the matrix exponential of A , because every eigenvalue λ of A corresponds to the eigenvalue $\exp(\lambda)$ of $\exp(A)$. In particular, given any logarithm of A , that is, any matrix L satisfying

$$\exp(L) = A$$

the determinant of A is given by

$$\det(A) = \exp(\text{tr}(L)).$$

For example, for $n = 2$, $n=3$, and $n = 4$, respectively,

$$\det(A) = ((\text{tr}A)^2 - \text{tr}(A^2))/2,$$

$$\det(A) = ((\text{tr}A)^3 - 3\text{tr}A \text{tr}(A^2) + 2\text{tr}(A^3))/6,$$

$$\det(A) = \left((\operatorname{tr}A)^4 - 6\operatorname{tr}(A^2)(\operatorname{tr}A)^2 + 3(\operatorname{tr}(A^2))^2 + 8\operatorname{tr}(A^3) \operatorname{tr}A - 6\operatorname{tr}(A^4) \right) / 24 .$$

cf. Cayley-Hamilton theorem. Such expressions are deducible from Newton's identities.

In the general case, ^[5]

$$\det(A) = \sum_{k_1, k_2, \dots, k_n} \prod_{l=1}^n \frac{(-1)^{k_l+1}}{l^{k_l} k_l!} \operatorname{tr}(A^l)^{k_l},$$

where the sum is taken over the set of all integers $k_l \geq 0$ satisfying the equation

$$\sum_{l=1}^n l k_l = n.$$

An arbitrary dimension n identity can be obtained from the Mercator series expansion of the logarithm,

$$\det(I + A) = \sum_{k=0}^{\infty} \frac{1}{k!} \left(- \sum_{j=1}^{\infty} \frac{(-1)^j}{j} \operatorname{tr}(A^j) \right)^k ,$$

where I is the identity matrix. The sum and the expansion of the exponential only need to go up to n instead of ∞ , since the determinant cannot exceed $O(A^n)$.

Cramer's rule

For a matrix equation

$$Ax = b$$

the solution is given by Cramer's rule:

$$x_i = \frac{\det(A_i)}{\det(A)} \quad i = 1, \dots, n$$

where A_i is the matrix formed by replacing the i th column of A by the column vector b . This follows immediately by column expansion of the determinant, i.e.

$$\det(A_i) = \det [a_1, \dots, b, \dots, a_n] = \sum_{j=1}^n x_j \det [a_1, \dots, a_{i-1}, a_j, a_{i+1}, \dots, a_n] = x_i \det(A)$$

where the vectors a_j are the columns of A . The rule is also implied by the identity

$$A \operatorname{adj}(A) = \operatorname{adj}(A) A = \det(A) I_n.$$

It has recently been shown that Cramer's rule can be implemented in $O(n^3)$ time,^[6] which is comparable to more common methods of solving systems of linear equations, such as LU, QR, or singular value decomposition.

Block matrices

Suppose $A, B, C,$ and D are matrices of dimension $(n \times n), (n \times m), (m \times n),$ and $(m \times m),$ respectively. Then

$$\det \begin{pmatrix} A & 0 \\ C & D \end{pmatrix} = \det \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} = \det(A) \det(D).$$

This can be seen from the Leibniz formula or by induction on n . When A is invertible, employing the following identity

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & 0 \\ C & I \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & D - CA^{-1}B \end{pmatrix}$$

leads to

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A) \det(D - CA^{-1}B).$$

When D is invertible, a similar identity with $\det(D)$ factored out can be derived analogously,^[7] that is,

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(D) \det(A - BD^{-1}C).$$

When the blocks are square matrices of the same order further formulas hold. For example, if C and D commute (i.e., $CD = DC$), then the following formula comparable to the determinant of a 2×2 matrix holds:^[8]

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(AD - BC).$$

When $A = D$ and $B = C$, the blocks are square matrices of the same order and the following formula holds (even if A and B do not commute)

$$\det \begin{pmatrix} A & B \\ B & A \end{pmatrix} = \det(A - B) \det(A + B).$$

When D is a 1×1 matrix, B is a column vector, and C is a row vector then

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = (D - 1) \det(A) + \det(A - BC) = (D + 1) \det A - \det(A + BC).$$

Derivative

By definition, e.g., using the Leibniz formula, the determinant of real (or analogously for complex) square matrices is a polynomial function from $\mathbf{R}^{n \times n}$ to \mathbf{R} . As such it is everywhere differentiable. Its derivative can be expressed using Jacobi's formula:

$$\frac{d \det(A)}{d\alpha} = \text{tr} \left(\text{adj}(A) \frac{dA}{d\alpha} \right).$$

where $\text{adj}(A)$ denotes the adjugate of A . In particular, if A is invertible, we have

$$\frac{d \det(A)}{d\alpha} = \det(A) \text{tr} \left(A^{-1} \frac{dA}{d\alpha} \right).$$

Expressed in terms of the entries of A , these are

$$\frac{\partial \det(A)}{\partial A_{ij}} = \text{adj}(A)_{ji} = \det(A) (A^{-1})_{ji}.$$

Yet another equivalent formulation is

$$\det(A + \epsilon X) - \det(A) = \text{tr}(\text{adj}(A)X)\epsilon + O(\epsilon^2) = \det(A) \text{tr}(A^{-1}X)\epsilon + O(\epsilon^2),$$

using big O notation. The special case where $A = I$, the identity matrix, yields

$$\det(I + \epsilon X) = 1 + \text{tr}(X)\epsilon + O(\epsilon^2).$$

This identity is used in describing the tangent space of certain matrix Lie groups.

If the matrix A is written as $A = [\mathbf{a} \ \mathbf{b} \ \mathbf{c}]$ where $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are vectors, then the gradient over one of the three vectors may be written as the cross product of the other two:

$$\begin{aligned} \nabla_{\mathbf{a}} \det(A) &= \mathbf{b} \times \mathbf{c} \\ \nabla_{\mathbf{b}} \det(A) &= \mathbf{c} \times \mathbf{a} \\ \nabla_{\mathbf{c}} \det(A) &= \mathbf{a} \times \mathbf{b}. \end{aligned}$$

Abstract algebraic aspects

Determinant of an endomorphism

The above identities concerning the determinant of products and inverses of matrices imply that similar matrices have the same determinant: two matrices A and B are similar, if there exists an invertible matrix X such that $A = X^{-1}BX$. Indeed, repeatedly applying the above identities yields

$$\det(A) = \det(X)^{-1} \det(BX) = \det(X)^{-1} \det(B) \det(X) = \det(B) \det(X)^{-1} \det(X) = \det(B).$$

The determinant is therefore also called a similarity invariant. The determinant of a linear transformation

$$T : V \rightarrow V$$

for some finite dimensional vector space V is defined to be the determinant of the matrix describing it, with respect to an arbitrary choice of basis in V . By the similarity invariance, this determinant is independent of the choice of the basis for V and therefore only depends on the endomorphism T .

Transformation on alternating multilinear n -forms

The vector space W of all alternating multilinear n -forms on an n -dimensional vector space V has dimension one. To each linear transformation T on V we associate a linear transformation T' on W , where for each w in W we define $(T'w)(x_1, \dots, x_n) = w(Tx_1, \dots, Tx_n)$. As a linear transformation on a one-dimensional space, T' is equivalent to a scalar multiple. We call this scalar the determinant of T .

Exterior algebra

The determinant can also be characterized as the unique function

$$D : M_n(K) \rightarrow K$$

from the set of all $n \times n$ matrices with entries in a field K to this field satisfying the following three properties: first, D is an n -linear function: considering all but one column of A fixed, the determinant is linear in the remaining column, that is

$$D(v_1, \dots, v_{i-1}, av_i + bw, v_{i+1}, \dots, v_n) = aD(v_1, \dots, v_{i-1}, v_i, v_{i+1}, \dots, v_n) + bD(v_1, \dots, v_{i-1}, w, v_{i+1}, \dots, v_n)$$

for any column vectors v_1, \dots, v_n , and w and any scalars (elements of K) a and b . Second, D is an alternating function: for any matrix A with two identical columns $D(A) = 0$. Finally, $D(I_n) = 1$. Here I_n is the identity matrix.

This fact also implies that every other n -linear alternating function $F: M_n(K) \rightarrow K$ satisfies

$$F(M) = F(I)D(M).$$

The last part in fact follows from the preceding statement: one easily sees that if F is nonzero it satisfies $F(I) \neq 0$, and function that associates $F(M)/F(I)$ to M satisfies all conditions of the theorem. The importance of stating this part is mainly that it remains valid^[9] if K is any commutative ring rather than a field, in which case the given argument does not apply.

The determinant of a linear transformation $A : V \rightarrow V$ of an n -dimensional vector space V can be formulated in a coordinate-free manner by considering the n th exterior power $\Lambda^n V$ of V . A induces a linear map

$$\begin{aligned} \Lambda^n A : \Lambda^n V &\rightarrow \Lambda^n V \\ v_1 \wedge v_2 \wedge \dots \wedge v_n &\mapsto Av_1 \wedge Av_2 \wedge \dots \wedge Av_n. \end{aligned}$$

As $\Lambda^n V$ is one-dimensional, the map $\Lambda^n A$ is given by multiplying with some scalar. This scalar coincides with the determinant of A , that is to say

$$(\Lambda^n A)(v_1 \wedge \dots \wedge v_n) = \det(A) \cdot v_1 \wedge \dots \wedge v_n.$$

This definition agrees with the more concrete coordinate-dependent definition. This follows from the characterization of the determinant given above. For example, switching two columns changes the parity of the

determinant; likewise, permuting the vectors in the exterior product $v_1 \wedge v_2 \wedge \dots \wedge v_n$ to $v_2 \wedge v_1 \wedge v_3 \wedge \dots \wedge v_n$, say, also alters the parity.

For this reason, the highest non-zero exterior power $\Lambda^n(V)$ is sometimes also called the determinant of V and similarly for more involved objects such as vector bundles or chain complexes of vector spaces. Minors of a matrix can also be cast in this setting, by considering lower alternating forms $\Lambda^k V$ with $k < n$.

Square matrices over commutative rings and abstract properties

The determinant of a matrix can be defined, for example using the Leibniz formula, for matrices with entries in any commutative ring. Briefly, a ring is a structure where addition, subtraction, and multiplication are defined. The commutativity requirement means that the product does not depend on the order of the two factors, i.e.,

$$r \cdot s = s \cdot r$$

is supposed to hold for all elements r and s of the ring. For example, the integers form a commutative ring.

ManyWikipedia:Please clarify of the above statements and notions carry over mutatis mutandis to determinants of these more general matrices: the determinant is multiplicative in this more general situation, and Cramer's rule also holds. A square matrix over a commutative ring R is invertible if and only if its determinant is a unit in R , that is, an element having a (multiplicative) inverse. (If R is a field, this latter condition is equivalent to the determinant being nonzero, thus giving back the above characterization.) For example, a matrix A with entries in \mathbf{Z} , the integers, is invertible (in the sense that the inverse matrix has again integer entries) if the determinant is $+1$ or -1 . Such a matrix is called unimodular.

The determinant defines a mapping

$$\text{GL}_n(R) \rightarrow R^\times,$$

between the group of invertible $n \times n$ matrices with entries in R and the multiplicative group of units in R . Since it respects the multiplication in both groups, this map is a group homomorphism. Secondly, given a ring homomorphism $f: R \rightarrow S$, there is a map $\text{GL}_n(R) \rightarrow \text{GL}_n(S)$ given by replacing all entries in R by their images under f . The determinant respects these maps, i.e., given a matrix $A = (a_{i,j})$ with entries in R , the identity

$$f(\det((a_{i,j}))) = \det((f(a_{i,j})))$$

holds. For example, the determinant of the complex conjugate of a complex matrix (which is also the determinant of its conjugate transpose) is the complex conjugate of its determinant, and for integer matrices: the reduction modulo m of the determinant of such a matrix is equal to the determinant of the matrix reduced modulo m (the latter determinant being computed using modular arithmetic). In the more high-brow parlance of category theory, the determinant is a natural transformation between the two functors GL_n and $(\cdot)^\times$. Adding yet another layer of abstraction, this is captured by saying that the determinant is a morphism of algebraic groups, from the general linear group to the multiplicative group,

$$\det : \text{GL}_n \rightarrow \mathbb{G}_m.$$

Generalizations and related notions

Infinite matrices

For matrices with an infinite number of rows and columns, the above definitions of the determinant do not carry over directly. For example, in Leibniz' formula, an infinite sum (all of whose terms are infinite products) would have to be calculated. Functional analysis provides different extensions of the determinant for such infinite-dimensional situations, which however only work for particular kinds of operators.

The Fredholm determinant defines the determinant for operators known as trace class operators by an appropriate generalization of the formula

$$\det(I + A) = \exp(\operatorname{tr}(\log(I + A))).$$

Another infinite-dimensional notion of determinant is the functional determinant.

Notions of determinant over non-commutative rings

For square matrices with entries in a non-commutative ring, there are various difficulties in defining determinants in a manner analogous to that for commutative rings. A meaning can be given to the Leibniz formula provided the order for the product is specified, and similarly for other ways to define the determinant, but non-commutativity then leads to the loss of many fundamental properties of the determinant, for instance the multiplicative property or the fact that the determinant is unchanged under transposition of the matrix. Over non-commutative rings, there is no reasonable notion of a multilinear form (if a bilinear form exists with a regular element of R as value on some pair of arguments, it can be used to show that all elements of R commute). Nevertheless various notions of non-commutative determinant have been formulated, which preserve some of the properties of determinants, notably quasideterminants and the Dieudonné determinant. It should also be noted that if one considers certain specific classes of matrices with non-commutative elements, then there are examples where one can define the determinant and prove linear algebra theorems which are very similar to their commutative analogs. Examples include: quantum groups and q -determinant; Capelli matrix and Capelli determinant; super-matrices and Berezinian; Manin matrices is the class of matrices which is most close to matrices with commutative elements.

Further variants

Determinants of matrices in superrings (that is, \mathbf{Z}_2 -graded rings) are known as Berezinians or superdeterminants.

The permanent of a matrix is defined as the determinant, except that the factors $\operatorname{sgn}(\sigma)$ occurring in Leibniz' rule are omitted. The immanant generalizes both by introducing a character of the symmetric group S_n in Leibniz' rule.

Calculation

Determinants are mainly used as a theoretical tool. They are rarely calculated explicitly in numerical linear algebra, where for applications like checking invertibility and finding eigenvalues the determinant has largely been supplanted by other techniques.^[10] Nonetheless, explicitly calculating determinants is required in some situations, and different methods are available to do so.

Naive methods of implementing an algorithm to compute the determinant include using Leibniz' formula or Laplace's formula. Both these approaches are extremely inefficient for large matrices, though, since the number of required operations grows very quickly: it is of order $n!$ (n factorial) for an $n \times n$ matrix M . For example, Leibniz' formula requires to calculate $n!$ products. Therefore, more involved techniques have been developed for calculating determinants.

Decomposition methods

Given a matrix A , some methods compute its determinant by writing A as a product of matrices whose determinants can be more easily computed. Such techniques are referred to as decomposition methods. Examples include the LU decomposition, the QR decomposition or the Cholesky decomposition (for positive definite matrices). These methods are of order $O(n^3)$, which is a significant improvement over $O(n!)$

The LU decomposition expresses A in terms of a lower triangular matrix L , an upper triangular matrix U and a permutation matrix P :

$$A = PLU.$$

The determinants of L and U can be quickly calculated, since they are the products of the respective diagonal entries. The determinant of P is just the sign ε of the corresponding permutation (which is $+1$ for an even number of permutations and is -1 for an uneven number of permutations). The determinant of A is then

$$\det(A) = \varepsilon \det(L) \cdot \det(U),$$

Moreover, the decomposition can be chosen such that L is a unitriangular matrix and therefore has determinant 1, in which case the formula further simplifies to

$$\det(A) = \varepsilon \det(U).$$

Further methods

If the determinant of A and the inverse of A have already been computed, the matrix determinant lemma allows to quickly calculate the determinant of $A + uv^T$, where u and v are column vectors.

Since the definition of the determinant does not need divisions, a question arises: do fast algorithms exist that do not need divisions? This is especially interesting for matrices over rings. Indeed algorithms with run-time proportional to n^4 exist. An algorithm of Mahajan and Vinay, and Berkowitz^[11] is based on closed ordered walks (short *clow*). It computes more products than the determinant definition requires, but some of these products cancel and the sum of these products can be computed more efficiently. The final algorithm looks very much like an iterated product of triangular matrices.

If two matrices of order n can be multiplied in time $M(n)$, where $M(n) \geq n^a$ for some $a > 2$, then the determinant can be computed in time $O(M(n))$.^[12] This means, for example, that an $O(n^{2.376})$ algorithm exists based on the Coppersmith–Winograd algorithm.

Algorithms can also be assessed according to their bit complexity, i.e., how many bits of accuracy are needed to store intermediate values occurring in the computation. For example, the Gaussian elimination (or LU decomposition) methods is of order $O(n^3)$, but the bit length of intermediate values can become exponentially long. The Bareiss Algorithm, on the other hand, is an exact-division method based on Sylvester's identity is also of order n^3 , but the bit complexity is roughly the bit size of the original entries in the matrix times n .

History

Historically, determinants were considered without reference to matrices: originally, a determinant was defined as a property of a system of linear equations. The determinant "determines" whether the system has a unique solution (which occurs precisely if the determinant is non-zero). In this sense, determinants were first used in the Chinese mathematics textbook *The Nine Chapters on the Mathematical Art* (九章算術, Chinese scholars, around the 3rd century BC). In Europe, 2×2 determinants were considered by Cardano at the end of the 16th century and larger ones by Leibniz.^{[13][14][15]}

In Europe, Cramer (1750) added to the theory, treating the subject in relation to sets of equations. The recurrence law was first announced by Bézout (1764).

It was Vandermonde (1771) who first recognized determinants as independent functions.^[1] Laplace (1772)^{[16][17]} gave the general method of expanding a determinant in terms of its complementary minors: Vandermonde had already given a special case. Immediately following, Lagrange (1773) treated determinants of the second and third order. Lagrange was the first to apply determinants to questions of elimination theory; he proved many special cases of general identities.

Gauss (1801) made the next advance. Like Lagrange, he made much use of determinants in the theory of numbers. He introduced the word *determinant* (Laplace had used *resultant*), though not in the present signification, but rather as applied to the discriminant of a quantic. Gauss also arrived at the notion of reciprocal (inverse) determinants, and came very near the multiplication theorem.

The next contributor of importance is Binet (1811, 1812), who formally stated the theorem relating to the product of two matrices of m columns and n rows, which for the special case of $m = n$ reduces to the multiplication theorem. On the same day (November 30, 1812) that Binet presented his paper to the Academy, Cauchy also presented one on the subject. (See Cauchy–Binet formula.) In this he used the word *determinant* in its present sense,^{[18][19]} summarized and simplified what was then known on the subject, improved the notation, and gave the multiplication theorem with a proof more satisfactory than Binet's.^[20] With him begins the theory in its generality.

The next important figure was Jacobi (from 1827). He early used the functional determinant which Sylvester later called the Jacobian, and in his memoirs in *Crelle* for 1841 he specially treats this subject, as well as the class of alternating functions which Sylvester has called *alternants*. About the time of Jacobi's last memoirs, Sylvester (1839) and Cayley began their work.^{[21][22]}

The study of special forms of determinants has been the natural result of the completion of the general theory. Axisymmetric determinants have been studied by Lebesgue, Hesse, and Sylvester; persymmetric determinants by Sylvester and Hankel; circulants by Catalan, Spottiswoode, Glaisher, and Scott; skew determinants and Pfaffians, in connection with the theory of orthogonal transformation, by Cayley; continuants by Sylvester; Wronskians (so called by Muir) by Christoffel and Frobenius; compound determinants by Sylvester, Reiss, and Picquet; Jacobians and Hessians by Sylvester; and symmetric gauche determinants by Trudi. Of the textbooks on the subject Spottiswoode's was the first. In America, Hanus (1886), Weld (1893), and Muir/Metzler (1933) published treatises.

Applications

Linear independence

As mentioned above, the determinant of a matrix (with real or complex entries, say) is zero if and only if the column vectors of the matrix are linearly dependent. Thus, determinants can be used to characterize linearly dependent vectors. For example, given two linearly independent vectors v_1, v_2 in \mathbf{R}^3 , a third vector v_3 lies in the plane spanned by the former two vectors exactly if the determinant of the 3×3 matrix consisting of the three vectors is zero. The same idea is also used in the theory of differential equations: given n functions $f_1(x), \dots, f_n(x)$ (supposed to be $n-1$ times differentiable), the Wronskian is defined to be

$$W(f_1, \dots, f_n)(x) = \begin{vmatrix} f_1(x) & f_2(x) & \cdots & f_n(x) \\ f_1'(x) & f_2'(x) & \cdots & f_n'(x) \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-1)}(x) & f_2^{(n-1)}(x) & \cdots & f_n^{(n-1)}(x) \end{vmatrix}.$$

It is non-zero (for some x) in a specified interval if and only if the given functions and all their derivatives up to order $n-1$ are linearly independent. If it can be shown that the Wronskian is zero everywhere on an interval then, in the case of analytic functions, this implies the given functions are linearly dependent. See the Wronskian and linear independence.

Orientation of a basis

The determinant can be thought of as assigning a number to every sequence of n vectors in \mathbf{R}^n , by using the square matrix whose columns are the given vectors. For instance, an orthogonal matrix with entries in \mathbf{R}^n represents an orthonormal basis in Euclidean space. The determinant of such a matrix determines whether the orientation of the basis is consistent with or opposite to the orientation of the standard basis. If the determinant is $+1$, the basis has the same orientation. If it is -1 , the basis has the opposite orientation.

More generally, if the determinant of A is positive, A represents an orientation-preserving linear transformation (if A is an orthogonal 2×2 or 3×3 matrix, this is a rotation), while if it is negative, A switches the orientation of the basis.

Volume and Jacobian determinant

As pointed out above, the absolute value of the determinant of real vectors is equal to the volume of the parallelepiped spanned by those vectors. As a consequence, if $f: \mathbf{R}^n \rightarrow \mathbf{R}^n$ is the linear map represented by the matrix A , and S is any measurable subset of \mathbf{R}^n , then the volume of $f(S)$ is given by $|\det(A)|$ times the volume of S . More generally, if the linear map $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is represented by the $m \times n$ matrix A , then the n -dimensional volume of $f(S)$ is given by:

$$\text{volume}(f(S)) = \sqrt{\det(A^T A)} \times \text{volume}(S).$$

By calculating the volume of the tetrahedron bounded by four points, they can be used to identify skew lines. The volume of any tetrahedron, given its vertices \mathbf{a} , \mathbf{b} , \mathbf{c} , and \mathbf{d} , is $(1/6) \cdot |\det(\mathbf{a} - \mathbf{b}, \mathbf{b} - \mathbf{c}, \mathbf{c} - \mathbf{d})|$, or any other combination of pairs of vertices that would form a spanning tree over the vertices.

For a general differentiable function, much of the above carries over by considering the Jacobian matrix of f . For

$$f : \mathbf{R}^n \rightarrow \mathbf{R}^n,$$

the Jacobian is the $n \times n$ matrix whose entries are given by

$$D(f) = \left(\frac{\partial f_i}{\partial x_j} \right)_{1 \leq i, j \leq n}.$$

Its determinant, the Jacobian determinant appears in the higher-dimensional version of integration by substitution: for suitable functions f and an open subset U of \mathbf{R}^n (the domain of f), the integral over $f(U)$ of some other function $\phi: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is given by

$$\int_{f(U)} \phi(\mathbf{v}) \, d\mathbf{v} = \int_U \phi(f(\mathbf{u})) |\det(Df)(\mathbf{u})| \, d\mathbf{u}.$$

The Jacobian also occurs in the inverse function theorem.

Vandermonde determinant (alternant)

Third order

$$\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \end{vmatrix} = (x_3 - x_2)(x_3 - x_1)(x_2 - x_1).$$

In general, the n th-order Vandermonde determinant is ^[23]

$$\begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & x_3^{n-1} & \cdots & x_n^{n-1} \end{vmatrix} = \prod_{1 \leq i < j \leq n} (x_j - x_i),$$

where the right-hand side is the continued product of all the differences that can be formed from the $n(n-1)/2$ pairs of numbers taken from x_1, x_2, \dots, x_n , with the order of the differences taken in the reversed order of the suffixes that are involved.

Circulants

Second order

$$\begin{vmatrix} x_1 & x_2 \\ x_2 & x_1 \end{vmatrix} = (x_1 + x_2)(x_1 - x_2).$$

Third order

$$\begin{vmatrix} x_1 & x_2 & x_3 \\ x_3 & x_1 & x_2 \\ x_2 & x_3 & x_1 \end{vmatrix} = (x_1 + x_2 + x_3)(x_1 + \omega x_2 + \omega^2 x_3)(x_1 + \omega^2 x_2 + \omega x_3),$$

where ω and ω^2 are the complex cube roots of 1. In general, the n th-order circulant determinant is

$$\begin{vmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ x_n & x_1 & x_2 & \cdots & x_{n-1} \\ x_{n-1} & x_n & x_1 & \cdots & x_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_2 & x_3 & x_4 & \cdots & x_1 \end{vmatrix} = \prod_{j=1}^n (x_1 + x_2 \omega_j + x_3 \omega_j^2 + \dots + x_n \omega_j^{n-1}),$$

where ω_j is an n th root of 1.

Notes

[1] Serge Lang, *Linear Algebra*, 2nd Edition, Addison-Wesley, 1971, pp 173, 191.

[2] WildLinAlg episode 4 (<http://www.youtube.com/watch?v=6XghF70fqkY>), Norman J Wildberger, Univ. of New South Wales, 2010, lecture via youtube

[3] In a non-commutative setting left-linearity (compatibility with left-multiplication by scalars) should be distinguished from right-linearity. Assuming linearity in the columns is taken to be left-linearity, one would have, for non-commuting scalars a, b :

UNIQ-math-0-efccba492c5c7a1c-QINU

a contradiction. There is no useful notion of multi-linear functions over a non-commutative ring.

[4] Proofs can be found in <http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/proof003.html>

[5] A proof can be found in the Appendix B of L. A. Kondratyuk, M. I. Krivoruchenko (1992), *Zeitschrift für Physik A* **344**, 99-115.

[6] Ken Habgood, Itamar Arel, *A condensation-based application of Cramer's rule for solving large-scale linear systems*, Journal of Discrete Algorithms, 10 (2012), pp. 98–109. Available online 1 July 2011, ISSN 1570–8667, 10.1016/j.jda.2011.06.007.

[7] These identities were taken from <http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/proof003.html>

[8] Proofs are given in J.R. Sylvester, Determinants of Block Matrices, Math. Gazette, 84 (2000), pp. 460–467, available at <http://www.jstor.org/stable/3620776>

[9] Roger Godement, *Cours d'Algèbre*, seconde édition, Hermann (1966), §23, Théorème 5, p. 303

[10] L. N. Trefethen and D. Bau, *Numerical Linear Algebra* (SIAM, 1997). e.g. in Lecture 1: "... we mention that the determinant, though a convenient notion theoretically, rarely finds a useful role in numerical algorithms."

[11] <http://page.inf.fu-berlin.de/~rote/Papers/pdf/Division-free+algorithms.pdf>

[12] J.R. Bunch and J.E. Hopcroft, Triangular factorization and inversion by fast matrix multiplication, *Mathematics of Computation*, 28 (1974) 231–236.

[13] Eves, H: "An Introduction to the History of Mathematics", pages 405, 493–494, Saunders College Publishing, 1990.

[14] A Brief History of Linear Algebra and Matrix Theory : <http://darkwing.uoregon.edu/~vitulli/441.sp04/LinAlgHistory.html>

[15] Cajori, F. *A History of Mathematics* p. 80 (<http://books.google.com/books?id=bBoPAAAAIAAJ&pg=PA80#v=onepage&f=false>)

[16] Expansion of determinants in terms of minors: Laplace, Pierre-Simon (de) "Recherches sur le calcul intégral et sur le système du monde," *Histoire de l'Académie Royale des Sciences* (Paris), seconde partie, pages 267–376 (1772).

[17] Muir, Sir Thomas, *The Theory of Determinants in the historical Order of Development* [London, England: Macmillan and Co., Ltd., 1906].

[18] The first use of the word "determinant" in the modern sense appeared in: Cauchy, Augustin-Louis "Memoire sur les fonctions qui ne peuvent obtenir que deux valeurs égales et des signes contraires par suite des transpositions operées entre les variables qu'elles renferment," which was first read at the Institute de France in Paris on November 30, 1812, and which was subsequently published in the *Journal de l'Ecole*

Polytechnique, Cahier 17, Tome 10, pages 29–112 (1815).

[19] Origins of mathematical terms: <http://jeff560.tripod.com/d.html>

[20] History of matrices and determinants: http://www-history.mcs.st-and.ac.uk/history/HistTopics/Matrices_and_determinants.html

[21] The first use of vertical lines to denote a determinant appeared in: Cayley, Arthur "On a theorem in the geometry of position," *Cambridge Mathematical Journal*, vol. 2, pages 267–271 (1841).

[22] History of matrix notation: <http://jeff560.tripod.com/matrices.html>

[23] Gradshteyn, I. S., I. M. Ryzhik: "Table of Integrals, Series, and Products", 14.31, Elsevier, 2007.

References

- Axler, Sheldon Jay (1997), *Linear Algebra Done Right* (2nd ed.), Springer-Verlag, ISBN 0-387-98259-0
- de Boer, Carl (1990), "An empty exercise" (<http://ftp.cs.wisc.edu/Approx/empty.pdf>), *ACM SIGNUM Newsletter* **25** (2): 3–7, doi: 10.1145/122272.122273 (<http://dx.doi.org/10.1145/122272.122273>).
- Lay, David C. (August 22, 2005), *Linear Algebra and Its Applications* (3rd ed.), Addison Wesley, ISBN 978-0-321-28713-7
- Meyer, Carl D. (February 15, 2001), *Matrix Analysis and Applied Linear Algebra* (<http://www.matrixanalysis.com/DownloadChapters.html>), Society for Industrial and Applied Mathematics (SIAM), ISBN 978-0-89871-454-8
- Muir, Thomas (1960) [1933], *A treatise on the theory of determinants*, Revised and enlarged by William H. Metzler, New York, NY: Dover
- Poole, David (2006), *Linear Algebra: A Modern Introduction* (2nd ed.), Brooks/Cole, ISBN 0-534-99845-3
- Anton, Howard (2005), *Elementary Linear Algebra (Applications Version)* (9th ed.), Wiley International
- Leon, Steven J. (2006), *Linear Algebra With Applications* (7th ed.), Pearson Prentice Hall

External links

- Hazewinkel, Michiel, ed. (2001), "Determinant" (<http://www.encyclopediaofmath.org/index.php?title=Determinant&oldid=12692>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Weisstein, Eric W., "Determinant" (<http://mathworld.wolfram.com/Determinant.html>), *MathWorld*.
- O'Connor, John J.; Robertson, Edmund F., "Matrices and determinants" (http://www-history.mcs.st-andrews.ac.uk/HistTopics/Matrices_and_determinants.html), *MacTutor History of Mathematics archive*, University of St Andrews.
- WebApp to calculate determinants and descriptively solve systems of linear equations (<http://sole.ooz.ie/en>)
- Determinant Interactive Program and Tutorial (<http://people.revoledu.com/kardi/tutorial/LinearAlgebra/MatrixDeterminant.html>)
- Online Matrix Calculator (<http://matrixcalc.org/en.index.html>)
- Linear algebra: determinants. (<http://www.umat.feec.vutbr.cz/~novakm/determinanty/en/>) Compute determinants of matrices up to order 6 using Laplace expansion you choose.
- Matrices and Linear Algebra on the Earliest Uses Pages (<http://www.economics.soton.ac.uk/staff/aldrich/matrices.htm>)
- Determinants explained in an easy fashion in the 4th chapter as a part of a Linear Algebra course. (<http://algebra.math.ust.hk/course/content.shtml>)
- Instructional Video on taking the determinant of an nxn matrix (Khan Academy) (<http://khanexercises.appspot.com/video?v=H9BWRVJNiv4>)
- Online matrix calculator (determinant, track, inverse, adjoint, transpose) (<http://www.elektro-energetika.cz/calculations/matreg.php?language=english>) Compute determinant of matrix up to order 8
- Derivation of Determinant of a Matrix (<http://www.amarketplaceofideas.com/math-derivation-of-matrix-determinant.htm>)

Minor (linear algebra)

In linear algebra, a **minor** of a matrix **A** is the determinant of some smaller square matrix, cut down from **A** by removing one or more of its rows or columns. Minors obtained by removing just one row and one column from square matrices (**first minors**) are required for calculating matrix **cofactors**, which in turn are useful for computing both the determinant and inverse of square matrices.

Definition and illustration

First minors

If **A** is a square matrix, then the **minor** of the entry in the *i*-th row and *j*-th column (also called the **(i,j) minor**, or a **first minor**^[1]) is the determinant of the submatrix formed by deleting the *i*-th row and *j*-th column. This number is often denoted $M_{i,j}$. The **(i,j) cofactor** is obtained by multiplying the minor by $(-1)^{i+j}$.

To illustrate these definitions, consider the following 3 by 3 matrix,

$$\begin{bmatrix} 1 & 4 & 7 \\ 3 & 0 & 5 \\ -1 & 9 & 11 \end{bmatrix}$$

To compute the minor $M_{2,3}$ and the cofactor $C_{2,3}$, we find the determinant of the above matrix with row 2 and column 3 removed.

$$M_{2,3} = \det \begin{bmatrix} 1 & 4 & \square \\ \square & \square & \square \\ -1 & 9 & \square \end{bmatrix} = \det \begin{bmatrix} 1 & 4 \\ -1 & 9 \end{bmatrix} = (9 - (-4)) = 13$$

So the cofactor of the (2,3) entry is

$$C_{23} = (-1)^{2+3}(M_{23}) = -13.$$

General definition

Let **A** be an $m \times n$ matrix and *k* an integer with $0 < k \leq m$, and $k \leq n$. A $k \times k$ **minor** of **A** is the determinant of a $k \times k$ matrix obtained from **A** by deleting $m - k$ rows and $n - k$ columns. For such a matrix there are a total of $\binom{m}{k} \cdot \binom{n}{k}$ minors of size $k \times k$.

Complement

The complement, $B_{ijk\dots pqr\dots}$, of a minor, $M_{ijk\dots pqr\dots}$, of a square matrix, **A**, is formed by the determinant of the matrix **A** from which all the rows (*ijk...*) and columns (*pqr...*) associated with $M_{ijk\dots pqr\dots}$ have been removed. The complement of the first minor of an element a_{ij} is merely that element.^[2]

Applications of minors and cofactors

Cofactor expansion of the determinant

The cofactors feature prominently in Laplace's formula for the expansion of determinants, which is a method of computing larger determinants in terms of smaller ones. Given the $n \times n$ matrix $(a_{i,j})$, the determinant of A (denoted $\det(A)$) can be written as the sum of the cofactors of any row or column of the matrix multiplied by the entries that generated them. In other words, the cofactor expansion along the j th column gives:

$$\det(\mathbf{A}) = a_{1j}C_{1j} + a_{2j}C_{2j} + a_{3j}C_{3j} + \dots + a_{nj}C_{nj} = \sum_{i=1}^n a_{ij}C_{ij}$$

The cofactor expansion along the i th row gives:

$$\det(\mathbf{A}) = a_{i1}C_{i1} + a_{i2}C_{i2} + a_{i3}C_{i3} + \dots + a_{in}C_{in} = \sum_{j=1}^n a_{ij}C_{ij}$$

Inverse of a matrix

One can write down the inverse of an invertible matrix by computing its cofactors by using Cramer's rule, as follows. The matrix formed by all of the cofactors of a square matrix \mathbf{A} is called the **cofactor matrix** (also called the **matrix of cofactors**):

$$\mathbf{C} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix}$$

Then the inverse of \mathbf{A} is the transpose of the cofactor matrix times the inverse of the determinant of A :

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \mathbf{C}^T.$$

The transpose of the cofactor matrix is called the adjugate matrix (also called the **classical adjoint**) of \mathbf{A} .

Other applications

Given an $m \times n$ matrix with real entries (or entries from any other field) and rank r , then there exists at least one non-zero $r \times r$ minor, while all larger minors are zero.

We will use the following notation for minors: if \mathbf{A} is an $m \times n$ matrix, I is a subset of $\{1, \dots, m\}$ with k elements and J is a subset of $\{1, \dots, n\}$ with k elements, then we write $[\mathbf{A}]_{I,J}$ for the $k \times k$ minor of \mathbf{A} that corresponds to the rows with index in I and the columns with index in J .

- If $I = J$, then $[\mathbf{A}]_{I,I}$ is called a **principal minor**.
- If the matrix that corresponds to a principal minor is a quadratic upper-left part of the larger matrix (i.e., it consists of matrix elements in rows and columns from 1 to k), then the principal minor is called a **leading principal minor**. For an $n \times n$ square matrix, there are n leading principal minors.
- For Hermitian matrices, the leading principal minors can be used to test for positive definiteness.

Both the formula for ordinary matrix multiplication and the Cauchy-Binet formula for the determinant of the product of two matrices are special cases of the following general statement about the minors of a product of two matrices. Suppose that \mathbf{A} is an $m \times n$ matrix, \mathbf{B} is an $n \times p$ matrix, I is a subset of $\{1, \dots, m\}$ with k elements and J is a subset of $\{1, \dots, p\}$ with k elements. Then

$$[\mathbf{AB}]_{I,J} = \sum_K [\mathbf{A}]_{I,K} [\mathbf{B}]_{K,J}$$

where the sum extends over all subsets K of $\{1, \dots, n\}$ with k elements. This formula is a straightforward extension of the Cauchy-Binet formula.

Multilinear algebra approach

A more systematic, algebraic treatment of the minor concept is given in multilinear algebra, using the wedge product: the k -minors of a matrix are the entries in the k th exterior power map.

If the columns of a matrix are wedged together k at a time, the $k \times k$ minors appear as the components of the resulting k -vectors. For example, the 2×2 minors of the matrix

$$\begin{pmatrix} 1 & 4 \\ 3 & -1 \\ 2 & 1 \end{pmatrix}$$

are -13 (from the first two rows), -7 (from the first and last row), and 5 (from the last two rows). Now consider the wedge product

$$(\mathbf{e}_1 + 3\mathbf{e}_2 + 2\mathbf{e}_3) \wedge (4\mathbf{e}_1 - \mathbf{e}_2 + \mathbf{e}_3)$$

where the two expressions correspond to the two columns of our matrix. Using the properties of the wedge product, namely that it is bilinear[3] and

$$\mathbf{e}_i \wedge \mathbf{e}_i = 0$$

and

$$\mathbf{e}_i \wedge \mathbf{e}_j = -\mathbf{e}_j \wedge \mathbf{e}_i,$$

we can simplify this expression to

$$-13\mathbf{e}_1 \wedge \mathbf{e}_2 - 7\mathbf{e}_1 \wedge \mathbf{e}_3 + 5\mathbf{e}_2 \wedge \mathbf{e}_3$$

where the coefficients agree with the minors computed earlier.

A remark about different notations

In some books ^[4] instead of *cofactor* the term *adjunct* is used. Moreover, it is denoted as \mathbf{A}_{ij} and defined in the same way as cofactor:

$$\mathbf{A}_{ij} = (-1)^{i+j} \mathbf{M}_{ij}$$

Using this notation the inverse matrix is written this way:

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix}$$

Keep in mind that *adjunct* is not adjugate or adjoint[3]. In modern terminology, the "adjoint" of a matrix most often refers to the corresponding adjoint operator.

References

- [1] Burnside, William Snow & Panton, Arthur William (1886) *Theory of Equations: with an Introduction to the Theory of Binary Algebraic Form* (http://books.google.com/books?id=BhgPAAAAIAAJ&pg=PA239&lpg=PA239&dq=first+minor+determinant&source=web&ots=BqWTIFMGIB&sig=aeCdnU1sARW9tshE_zhirJZ5dRU&hl=en).
- [2] Bertha Jeffreys, *Methods of Mathematical Physics* (http://books.google.co.uk/books?id=Qs-xdYBQ_5wC&pg=PA135), p.135, Cambridge University Press, 1999 ISBN 0-521-66402-0.
- [3] [http://toolsserver.org/%7Edispenser/cgi-bin/dab_solver.py?page=Minor_\(linear_algebra\)&editintro=Template:Disambiguation_needed/editintro&client=Template:Dn](http://toolsserver.org/%7Edispenser/cgi-bin/dab_solver.py?page=Minor_(linear_algebra)&editintro=Template:Disambiguation_needed/editintro&client=Template:Dn)
- [4] Felix Gantmacher, *Theory of matrices* (1st ed., original language is Russian), Moscow: State Publishing House of technical and theoretical literature, 1953, p.491,

External links

- MIT Linear Algebra Lecture on Cofactors (<http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/lecture-19-determinant-formulas-and-cofactors/>) at Google Video, from MIT OpenCourseWare
- PlanetMath entry of *Cofactors* (<http://planetmath.org/encyclopedia/Cofactor.html>)
- Springer Encyclopedia of Mathematics entry for *Minor* (<http://www.encyclopediaofmath.org/index.php/Minor>)

Adjugate matrix

In linear algebra, the **adjugate** or **classical adjoint** (occasionally referred to as **adjunct**) of a square matrix is the transpose of the cofactor matrix.

The adjugate has sometimes been called the "adjoint", but today the "adjoint" of a matrix normally refers to its corresponding adjoint operator, which is its conjugate transpose.

Definition

The adjugate of A is the transpose of the cofactor matrix C of A :

$$\operatorname{adj}(\mathbf{A}) = \mathbf{C}^T.$$

In more detail: suppose R is a commutative ring and \mathbf{A} is an $n \times n$ matrix with entries from R .

- The (i,j) *minor* of \mathbf{A} , denoted \mathbf{A}_{ij} , is the determinant of the $(n-1) \times (n-1)$ matrix that results from deleting row i and column j of \mathbf{A} .
- The cofactor matrix of \mathbf{A} is the $n \times n$ matrix \mathbf{C} whose (i,j) entry is the (i,j) *cofactor* of \mathbf{A} :

$$\mathbf{C}_{ij} = (-1)^{i+j} \mathbf{A}_{ij}.$$

- The adjugate of \mathbf{A} is the transpose of \mathbf{C} , that is, the $n \times n$ matrix whose (i,j) entry is the (j,i) cofactor of \mathbf{A} :

$$\operatorname{adj}(\mathbf{A})_{ij} = \mathbf{C}_{ji}.$$

The adjugate is defined as it is so that the product of A and its adjugate yields a diagonal matrix whose diagonal entries are $\det(\mathbf{A})$:

$$\mathbf{A} \operatorname{adj}(\mathbf{A}) = \det(\mathbf{A}) \mathbf{I}.$$

\mathbf{A} is invertible if and only if $\det(\mathbf{A})$ is an invertible element of R , and in that case the equation above yields:

$$\begin{aligned} \operatorname{adj}(\mathbf{A}) &= \det(\mathbf{A}) \mathbf{A}^{-1}, \\ \mathbf{A}^{-1} &= \frac{1}{\det(\mathbf{A})} \operatorname{adj}(\mathbf{A}). \end{aligned}$$

Examples

2 × 2 generic matrix

The adjugate of the 2 × 2 matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is

$$\text{adj}(\mathbf{A}) = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

It is seen that $\det(\text{adj}(\mathbf{A})) = \det(\mathbf{A})$ and $\text{adj}(\text{adj}(\mathbf{A})) = \mathbf{A}$.

3 × 3 generic matrix

Consider the 3 × 3 matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

Its adjugate is the transpose of the cofactor matrix

$$\mathbf{C} = \begin{pmatrix} + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & - \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} & + \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\ - \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} & + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & - \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} \\ + \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} & - \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} & + \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{pmatrix} = \begin{pmatrix} + \begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} & - \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} & + \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix} \\ - \begin{vmatrix} 2 & 3 \\ 8 & 9 \end{vmatrix} & + \begin{vmatrix} 1 & 3 \\ 7 & 9 \end{vmatrix} & - \begin{vmatrix} 1 & 2 \\ 7 & 8 \end{vmatrix} \\ + \begin{vmatrix} 2 & 3 \\ 5 & 6 \end{vmatrix} & - \begin{vmatrix} 1 & 3 \\ 4 & 6 \end{vmatrix} & + \begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix} \end{pmatrix}$$

So that we have

$$\text{adj}(\mathbf{A}) = \begin{pmatrix} + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} & - \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} & + \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \\ - \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} & + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} & - \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} \\ + \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} & - \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} & + \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \end{pmatrix} = \begin{pmatrix} + \begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} & - \begin{vmatrix} 2 & 3 \\ 8 & 9 \end{vmatrix} & + \begin{vmatrix} 2 & 3 \\ 5 & 6 \end{vmatrix} \\ - \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} & + \begin{vmatrix} 1 & 3 \\ 7 & 9 \end{vmatrix} & - \begin{vmatrix} 1 & 3 \\ 4 & 6 \end{vmatrix} \\ + \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix} & - \begin{vmatrix} 1 & 2 \\ 7 & 8 \end{vmatrix} & + \begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix} \end{pmatrix}$$

where

$$\begin{vmatrix} a_{im} & a_{in} \\ a_{jm} & a_{jn} \end{vmatrix} = \det \begin{pmatrix} a_{im} & a_{in} \\ a_{jm} & a_{jn} \end{pmatrix}.$$

Therefore \mathbf{C} , the matrix of cofactors for \mathbf{A} , is

$$\mathbf{C} = \begin{pmatrix} -3 & 6 & -3 \\ 6 & -12 & 6 \\ -3 & 6 & -3 \end{pmatrix}$$

The adjugate is the *transpose* of the cofactor matrix. Thus, for instance, the (3,2) entry of the adjugate is the (2,3) cofactor of \mathbf{A} . (In this example, \mathbf{C} happens to be its own transpose, so $\text{adj}(\mathbf{A}) = \mathbf{C}$.)

3 × 3 numeric matrix

As a specific example, we have

$$\text{adj} \begin{pmatrix} -3 & 2 & -5 \\ -1 & 0 & -2 \\ 3 & -4 & 1 \end{pmatrix} = \begin{pmatrix} -8 & 18 & -4 \\ -5 & 12 & -1 \\ 4 & -6 & 2 \end{pmatrix}.$$

The -6 in the third row, second column of the adjugate was computed as follows:

$$(-1)^{2+3} \det \begin{pmatrix} -3 & 2 \\ 3 & -4 \end{pmatrix} = -((-3)(-4) - (3)(2)) = -6.$$

Again, the (3,2) entry of the adjugate is the (2,3) cofactor of A . Thus, the submatrix

$$\begin{pmatrix} -3 & 2 \\ 3 & -4 \end{pmatrix}$$

was obtained by deleting the second row and third column of the original matrix A .

Properties

The adjugate has the properties

$$\begin{aligned} \text{adj}(\mathbf{I}) &= \mathbf{I}, \\ \text{adj}(\mathbf{AB}) &= \text{adj}(\mathbf{B}) \text{adj}(\mathbf{A}), \\ \text{adj}(c\mathbf{A}) &= c^{n-1} \text{adj}(\mathbf{A}) \end{aligned}$$

for $n \times n$ matrices A and B . The second line follows from equations $\text{adj}(\mathbf{B})\text{adj}(\mathbf{A}) = \det(\mathbf{B})\mathbf{B}^{-1} \det(\mathbf{A})\mathbf{A}^{-1} = \det(\mathbf{AB})(\mathbf{AB})^{-1}$. Substituting in the second line $\mathbf{B} = \mathbf{A}^{m-1}$ and performing the recursion, one gets for all integer m

$$\text{adj}(\mathbf{A}^m) = \text{adj}(\mathbf{A})^m.$$

The adjugate preserves transposition:

$$\text{adj}(\mathbf{A}^\top) = \text{adj}(\mathbf{A})^\top.$$

Furthermore,

$$\begin{aligned} \det(\text{adj}(\mathbf{A})) &= \det(\mathbf{A})^{n-1}, \\ \text{adj}(\text{adj}(\mathbf{A})) &= \det(\mathbf{A})^{n-2} \mathbf{A} \end{aligned}$$

and, if $\det(\mathbf{A})$ is a unit, then $\det(\text{adj}(\mathbf{A})) = \det(\mathbf{A})$ and $\text{adj}(\text{adj}(\mathbf{A})) = \mathbf{A}$.

Inverses

As a consequence of Laplace's formula for the determinant of an $n \times n$ matrix A , we have

$$\mathbf{A} \text{adj}(\mathbf{A}) = \text{adj}(\mathbf{A}) \mathbf{A} = \det(\mathbf{A}) \mathbf{I}_n \quad (*)$$

where \mathbf{I}_n is the $n \times n$ identity matrix. Indeed, the (i,i) entry of the product $\mathbf{A} \text{adj}(\mathbf{A})$ is the scalar product of row i of A with row i of the cofactor matrix C , which is simply the Laplace formula for $\det(A)$ expanded by row i . Moreover, for $i \neq j$ the (i,j) entry of the product is the scalar product of row i of A with row j of C , which is the Laplace formula for the determinant of a matrix whose i and j rows are equal and is therefore zero.

From this formula follows one of the most important results in matrix algebra: A matrix A over a commutative ring R is invertible if and only if $\det(A)$ is invertible in R .

For if A is an invertible matrix then

$$1 = \det(\mathbf{I}_n) = \det(\mathbf{AA}^{-1}) = \det(\mathbf{A}) \det(\mathbf{A}^{-1}),$$

and equation (*) above shows that

$$\mathbf{A}^{-1} = \det(\mathbf{A})^{-1} \text{adj}(\mathbf{A}).$$

See also Cramer's rule.

Characteristic polynomial

If $p(t) = \det(\mathbf{A} - t \mathbf{I})$ is the characteristic polynomial of \mathbf{A} and we define the polynomial $q(t) = (p(0) - p(t))/t$, then

$$\text{adj}(\mathbf{A}) = q(\mathbf{A}) = -(p_1 \mathbf{I} + p_2 \mathbf{A} + p_3 \mathbf{A}^2 + \cdots + p_n \mathbf{A}^{n-1}),$$

where p_j are the coefficients of $p(t)$,

$$p(t) = p_0 + p_1 t + p_2 t^2 + \cdots + p_n t^n.$$

Jacobi's formula

The adjugate also appears in Jacobi's formula for the derivative of the determinant:

$$\frac{d}{d\alpha} \det(A) = \text{tr} \left(\text{adj}(A) \frac{dA}{d\alpha} \right).$$

References

- Strang, Gilbert (1988). "Section 4.4: Applications of determinants". *Linear Algebra and its Applications* (3rd ed.). Harcourt Brace Jovanovich. pp. 231–232. ISBN 0-15-551005-3.

External links

- Matrix Reference Manual (<http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/property.html#adjoint>)
- Online matrix calculator (determinant, track, inverse, adjoint, transpose) (<http://www.elektro-energetika.cz/calculations/matreg.php?language=english>) Compute Adjugate matrix up to order 8
- "adjugate of { { a, b, c }, { d, e, f }, { g, h, i } }" ([http://www.wolframalpha.com/input/?i=adjugate+of+\({+{+a,+b,+c+},{+d,+e,+f+},{+g,+h,+i+}\)](http://www.wolframalpha.com/input/?i=adjugate+of+({+{+a,+b,+c+},{+d,+e,+f+},{+g,+h,+i+}))). *Wolfram Alpha*.

Invertible matrix

In linear algebra an n -by- n (square) matrix \mathbf{A} is called **invertible** (some authors use **nonsingular** or **nondegenerate**) if there exists an n -by- n matrix \mathbf{B} such that

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$$

where \mathbf{I}_n denotes the n -by- n identity matrix and the multiplication used is ordinary matrix multiplication. If this is the case, then the matrix \mathbf{B} is uniquely determined by \mathbf{A} and is called the **inverse** of \mathbf{A} , denoted by \mathbf{A}^{-1} . It follows from the theory of matrices that if

$$\mathbf{AB} = \mathbf{I}$$

for *finite square* matrices \mathbf{A} and \mathbf{B} , then also

$$\mathbf{BA} = \mathbf{I}.$$

Non-square matrices (m -by- n matrices for which $m \neq n$) do not have an inverse. However, in some cases such a matrix may have a left inverse or right inverse. If \mathbf{A} is m -by- n and the rank of \mathbf{A} is equal to n , then \mathbf{A} has a left inverse: an n -by- m matrix \mathbf{B} such that $\mathbf{BA} = \mathbf{I}$. If \mathbf{A} has rank m , then it has a right inverse: an n -by- m matrix \mathbf{B} such that $\mathbf{AB} = \mathbf{I}$.

A square matrix that is not invertible is called **singular** or **degenerate**. A square matrix is singular if and only if its determinant is 0. Singular matrices are rare in the sense that a square matrix randomly selected from a continuous uniform distribution on its entries will almost never be singular.

While the most common case is that of matrices over the real or complex numbers, all these definitions can be given for matrices over any commutative ring. However, in this case the condition for a square matrix to be invertible is that its determinant is invertible in the ring, which in general is a much stricter requirement than being nonzero. The conditions for existence of left-inverse resp. right-inverse are more complicated since a notion of rank does not exist over rings.

Matrix inversion is the process of finding the matrix \mathbf{B} that satisfies the prior equation for a given invertible matrix \mathbf{A} .

Properties

The invertible matrix theorem

Let \mathbf{A} be a square n by n matrix over a field K (for example the field \mathbf{R} of real numbers). The following statements are equivalent:

\mathbf{A} is invertible, i.e. \mathbf{A} has an inverse, is nonsingular, or is nondegenerate.

\mathbf{A} is row-equivalent to the n -by- n identity matrix \mathbf{I}_n .

\mathbf{A} is column-equivalent to the n -by- n identity matrix \mathbf{I}_n .

\mathbf{A} has n pivot positions.

$\det \mathbf{A} \neq 0$. In general, a square matrix over a commutative ring is invertible if and only if its determinant is a unit in that ring.

\mathbf{A} has full rank; that is, $\text{rank } \mathbf{A} = n$.

The equation $\mathbf{Ax} = \mathbf{0}$ has only the trivial solution $\mathbf{x} = \mathbf{0}$

Null $\mathbf{A} = \{0\}$

The equation $\mathbf{Ax} = \mathbf{b}$ has exactly one solution for each \mathbf{b} in K^n .

The columns of \mathbf{A} are linearly independent.

The columns of \mathbf{A} span K^n

$\text{Col } \mathbf{A} = K^n$

The columns of \mathbf{A} form a basis of K^n .

The linear transformation mapping \mathbf{x} to \mathbf{Ax} is a bijection from K^n to K^n .

There is an n by n matrix \mathbf{B} such that $\mathbf{AB} = \mathbf{I}_n = \mathbf{BA}$.

The transpose \mathbf{A}^T is an invertible matrix (hence rows of \mathbf{A} are linearly independent, span K^n , and form a basis of K^n).

The number 0 is not an eigenvalue of \mathbf{A} .

The matrix \mathbf{A} can be expressed as a finite product of elementary matrices.

The matrix \mathbf{A} has a left inverse (i.e. there exists a \mathbf{B} such that $\mathbf{BA} = \mathbf{I}$) or a right inverse (i.e. there exists a \mathbf{C} such that $\mathbf{AC} = \mathbf{I}$), in which case both left and right inverses exist and $\mathbf{B} = \mathbf{C} = \mathbf{A}^{-1}$.

Other properties

Furthermore, the following properties hold for an invertible matrix \mathbf{A} :

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$;
- $(k\mathbf{A})^{-1} = k^{-1}\mathbf{A}^{-1}$ for nonzero scalar k ;
- $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$;
- For any invertible n -by- n matrices \mathbf{A} and \mathbf{B} , $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. More generally, if $\mathbf{A}_1, \dots, \mathbf{A}_k$ are invertible n -by- n matrices, then $(\mathbf{A}_1\mathbf{A}_2 \dots \mathbf{A}_{k-1}\mathbf{A}_k)^{-1} = \mathbf{A}_k^{-1}\mathbf{A}_{k-1}^{-1} \dots \mathbf{A}_2^{-1}\mathbf{A}_1^{-1}$;
- $\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1}$.

A matrix that is its own inverse, i.e. $\mathbf{A} = \mathbf{A}^{-1}$ and $\mathbf{A}^2 = \mathbf{I}$, is called an involution.

Density

Over the field of real numbers, the set of singular n -by- n matrices, considered as a subset of $\mathbf{R}^{n \times n}$, is a null set, i.e., has Lebesgue measure zero. This is true because singular matrices are the roots of the polynomial function in the entries of the matrix given by the determinant. Thus in the language of measure theory, almost all n -by- n matrices are invertible.

Furthermore the n -by- n invertible matrices are a dense open set in the topological space of all n -by- n matrices. Equivalently, the set of singular matrices is closed and nowhere dense in the space of n -by- n matrices.

In practice however, one may encounter non-invertible matrices. And in numerical calculations, matrices which are invertible, but close to a non-invertible matrix, can still be problematic; such matrices are said to be ill-conditioned.

Methods of matrix inversion

Gaussian elimination

Gauss–Jordan elimination is an algorithm that can be used to determine whether a given matrix is invertible and to find the inverse. An alternative is the LU decomposition which generates upper and lower triangular matrices which are easier to invert.

Newton's method

A generalisation of Newton's method as used for a multiplicative inverse algorithm may be convenient, if it is convenient to find a suitable starting seed:

$$X_{k+1} = 2X_k - X_kAX_k.$$

Victor Pan and John Reif have done work that includes ways of generating a starting seed. Otherwise, the method may be adapted to use the starting seed from a trivial starting case by using a homotopy to "walk" in small steps from that to the matrix needed, "dragging" the inverses with them:

$$X_{k+1} = 2X_k - X_k A_{k+1} X_k, \text{ where } A_0 = S, X_0 = S^{-1}, \text{ and } A_N = A \text{ for some terminating } N, \text{ perhaps followed by another few iterations at } A \text{ to settle the inverse.}$$

Using this simplistically on real valued matrices would lead the homotopy through a degenerate matrix about half the time, so complex valued matrices should be used to bypass that, e.g. by using a starting seed S that has i in the first entry, 1 on the rest of the leading diagonal, and 0 elsewhere. If complex arithmetic is not directly available, it may be emulated at a small cost in computer memory by replacing each complex matrix element $a+bi$ with a 2×2 real valued submatrix of the form $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$ (see square root of a matrix).

Newton's method is particularly useful when dealing with families of related matrices that behave enough like the sequence manufactured for the homotopy above: sometimes a good starting point for refining an approximation for the new inverse can be the already obtained inverse of a previous matrix that nearly matches the current matrix, e.g. the pair of sequences of inverse matrices used in obtaining matrix square roots by Denman-Beavers iteration; this may need more than one pass of the iteration at each new matrix, if they are not close enough together for just one to be enough. Newton's method is also useful for "touch up" corrections to the Gauss–Jordan algorithm which has been contaminated by small errors due to imperfect computer arithmetic.

Cayley–Hamilton method

Cayley–Hamilton theorem allows to represent the inverse of \mathbf{A} in terms of $\det(\mathbf{A})$, traces and powers of \mathbf{A}

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \sum_{s=0}^{n-1} \mathbf{A}^s \sum_{k_1, k_2, \dots, k_{n-1}} \prod_{l=1}^{n-1} \frac{(-1)^{k_l+1}}{l^{k_l} k_l!} \text{tr}(\mathbf{A}^l)^{k_l},$$

where n is dimension of \mathbf{A} , and the sum is taken over s and the sets of all $k_l \geq 0$ satisfying the linear Diophantine equation

$$s + \sum_{l=1}^{n-1} l k_l = n - 1.$$

Eigendecomposition

If matrix \mathbf{A} can be eigendecomposed and if none of its eigenvalues are zero, then \mathbf{A} is nonsingular and its inverse is given by

$$\mathbf{A}^{-1} = \mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^{-1}$$

where \mathbf{Q} is the square ($N \times N$) matrix whose i^{th} column is the eigenvector \mathbf{q}_i of \mathbf{A} and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, i.e., $\Lambda_{ii} = \lambda_i$. Furthermore, because $\mathbf{\Lambda}$ is a diagonal matrix, its inverse is easy to calculate:

$$[\mathbf{\Lambda}^{-1}]_{ii} = \frac{1}{\lambda_i}$$

Cholesky decomposition

If matrix \mathbf{A} is positive definite, then its inverse can be obtained as

$$\mathbf{A}^{-1} = (\mathbf{L}^*)^{-1}\mathbf{L}^{-1},$$

where \mathbf{L} is the lower triangular Cholesky decomposition of \mathbf{A} , and \mathbf{L}^* denotes the conjugate transpose of \mathbf{L} .

Analytic solution

Writing the transpose of the matrix of cofactors, known as an adjugate matrix, can also be an efficient way to calculate the inverse of *small* matrices, but this recursive method is inefficient for large matrices. To determine the inverse, we calculate a matrix of cofactors:

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\mathbf{C}^T = \frac{1}{|\mathbf{A}|} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{21} & \cdots & \mathbf{C}_{n1} \\ \mathbf{C}_{12} & \mathbf{C}_{22} & \cdots & \mathbf{C}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{1n} & \mathbf{C}_{2n} & \cdots & \mathbf{C}_{nn} \end{pmatrix}$$

so that

$$(\mathbf{A}^{-1})_{ij} = \frac{1}{|\mathbf{A}|} (\mathbf{C}^T)_{ij} = \frac{1}{|\mathbf{A}|} (\mathbf{C}_{ji})$$

where $|\mathbf{A}|$ is the determinant of \mathbf{A} , \mathbf{C} is the matrix of cofactors, and \mathbf{C}^T represents the matrix transpose.

Inversion of 2x2 matrices

The *cofactor equation* listed above yields the following result for 2x2 matrices. Inversion of these matrices can be done easily as follows:^[1]

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

This is possible because $1/(ad-bc)$ is the reciprocal of the determinant of the matrix in question, and the same strategy could be used for other matrix sizes.

The Cayley–Hamilton method gives

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} [\text{tr}\mathbf{A} - \mathbf{A}].$$

Inversion of 3x3 matrices

A computationally efficient 3x3 matrix inversion is given by

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} A & B & C \\ D & E & F \\ G & H & I \end{bmatrix}^T = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} A & D & G \\ B & E & H \\ C & F & I \end{bmatrix}$$

where the determinant of \mathbf{A} can be computed by applying the rule of Sarrus as follows:

$$\det(\mathbf{A}) = a(ei - fh) - b(id - fg) + c(dh - eg).$$

If the determinant is non-zero, the matrix is invertible, with the elements of the above matrix on the right side given by

$$\begin{aligned} A &= (ei - fh) & D &= -(bi - ch) & G &= (bf - ce) \\ B &= -(di - fg) & E &= (ai - cg) & H &= -(af - cd) \\ C &= (dh - eg) & F &= -(ah - bg) & I &= (ae - bd) \end{aligned}$$

The Cayley–Hamilton decomposition gives

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \left[\frac{1}{2} ((\text{tr}\mathbf{A})^2 - \text{tr}\mathbf{A}^2) - \mathbf{A}\text{tr}\mathbf{A} + \mathbf{A}^2 \right].$$

The general 3x3 inverse can be expressed concisely in terms of the cross product and triple product:

If a matrix $\mathbf{A} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2]$ (consisting of three column vectors, \mathbf{x}_0 , \mathbf{x}_1 , and \mathbf{x}_2) is invertible, its inverse is given by

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} (\mathbf{x}_1 \times \mathbf{x}_2)^T \\ (\mathbf{x}_2 \times \mathbf{x}_0)^T \\ (\mathbf{x}_0 \times \mathbf{x}_1)^T \end{bmatrix}.$$

Note that $\det(\mathbf{A})$ is equal to the triple product of \mathbf{x}_0 , \mathbf{x}_1 , and \mathbf{x}_2 —the volume of the parallelepiped formed by the rows or columns:

$$\det(\mathbf{A}) = \mathbf{x}_0 \cdot (\mathbf{x}_1 \times \mathbf{x}_2).$$

The correctness of the formula can be checked by using cross- and triple-product properties and by noting that for groups, left and right inverses always coincide. Intuitively, because of the cross products, each row of \mathbf{A}^{-1} is orthogonal to the non-corresponding two columns of \mathbf{A} (causing the off-diagonal terms of $\mathbf{I} = \mathbf{A}^{-1}\mathbf{A}$ be zero).

Dividing by

$$\det(\mathbf{A}) = \mathbf{x}_0 \cdot (\mathbf{x}_1 \times \mathbf{x}_2)$$

causes the diagonal elements of $\mathbf{I} = \mathbf{A}^{-1}\mathbf{A}$ to be unity. For example, the first diagonal is:

$$1 = \frac{1}{\mathbf{x}_0 \cdot (\mathbf{x}_1 \times \mathbf{x}_2)} \mathbf{x}_0 \cdot (\mathbf{x}_1 \times \mathbf{x}_2).$$

Inversion of 4x4 matrices

With increasing dimension, expressions for the inverse of \mathbf{A} get complicated. For $n = 4$ the Cayley-Hamilton method leads to an expression that is still tractable:

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \left[\frac{1}{6} ((\text{tr}\mathbf{A})^3 - 3\text{tr}\mathbf{A}\text{tr}\mathbf{A}^2 + 2\text{tr}\mathbf{A}^3) - \frac{1}{2}\mathbf{A} ((\text{tr}\mathbf{A})^2 - \text{tr}\mathbf{A}^2) + \mathbf{A}^2\text{tr}\mathbf{A} - \mathbf{A}^3 \right].$$

Blockwise inversion

Matrices can also be *inverted blockwise* by using the following analytic inversion formula:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix} \tag{1}$$

where \mathbf{A} , \mathbf{B} , \mathbf{C} and \mathbf{D} are matrix sub-blocks of arbitrary size. (\mathbf{A} and \mathbf{D} must be square, so that they can be inverted. Furthermore, \mathbf{A} and $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ must be nonsingular.) This strategy is particularly advantageous if \mathbf{A} is diagonal and $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ (the Schur complement of \mathbf{A}) is a small matrix, since they are the only matrices requiring inversion. This technique was reinvented several times and is due to Hans Boltz (1923),^[citation needed] who used it for the inversion of geodetic matrices, and Tadeusz Banachiewicz (1937), who generalized it and proved its correctness.

The nullity theorem says that the nullity of \mathbf{A} equals the nullity of the sub-block in the lower right of the inverse matrix, and that the nullity of \mathbf{B} equals the nullity of the sub-block in the upper right of the inverse matrix.

The inversion procedure that led to Equation (1) performed matrix block operations that operated on \mathbf{C} and \mathbf{D} first. Instead, if \mathbf{A} and \mathbf{B} are operated on first, and provided \mathbf{D} and $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ are nonsingular, the result is

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}. \tag{2}$$

Equating Equations (1) and (2) leads to

$$(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{3}$$

$$\begin{aligned} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} &= \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} &= (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} \\ \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} &= (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{aligned}$$

where Equation (3) is the matrix inversion lemma, which is equivalent to the binomial inverse theorem.

Since a blockwise inversion of an $n \times n$ matrix requires inversion of two half-sized matrices and 6 multiplications between two half-sized matrices, it can be shown that a divide and conquer algorithm that uses blockwise inversion to invert a matrix runs with the same time complexity as the matrix multiplication algorithm that is used internally.^[2] There exist matrix multiplication algorithms with a complexity of $O(n^{2.3727})$ operations, while the best proven lower bound is $\Omega(n^2 \log n)$.^[3]

By Neumann series

If a matrix \mathbf{A} has the property that

$$\lim_{n \rightarrow \infty} (\mathbf{I} - \mathbf{A})^n = \mathbf{0}$$

then \mathbf{A} is nonsingular and its inverse may be expressed by a Neumann series:

$$\mathbf{A}^{-1} = \sum_{n=0}^{\infty} (\mathbf{I} - \mathbf{A})^n.$$

Truncating the sum results in an "approximate" inverse which may be useful as a preconditioner. Note that a truncated series can be accelerated exponentially by noting that the Neumann series is a geometric sum. Therefore, if one wishes to compute 2^L terms, one merely need the moments $\mathbf{A}, \mathbf{A}^2, \mathbf{A}^4, \dots, \mathbf{A}^{2^L}$ which can be found through L matrix multiplications. Then another L matrix multiplications are needed to obtains the final result by multiplying all the moments together. Therefore, $2L$ matrix multiplications are needed to compute 2^L terms of the sum.

More generally, if \mathbf{A} is "near" the invertible matrix \mathbf{X} in the sense that

$$\lim_{n \rightarrow \infty} (\mathbf{I} - \mathbf{X}^{-1}\mathbf{A})^n = \mathbf{0} \quad \text{or} \quad \lim_{n \rightarrow \infty} (\mathbf{I} - \mathbf{A}\mathbf{X}^{-1})^n = \mathbf{0}$$

then \mathbf{A} is nonsingular and its inverse is

$$\mathbf{A}^{-1} = \sum_{n=0}^{\infty} (\mathbf{X}^{-1}(\mathbf{A} - \mathbf{X}))^n \mathbf{X}^{-1}.$$

If it is also the case that $\mathbf{A} - \mathbf{X}$ has rank 1 then this simplifies to

$$\mathbf{A}^{-1} = \mathbf{X}^{-1} - \frac{\mathbf{X}^{-1}(\mathbf{A} - \mathbf{X})\mathbf{X}^{-1}}{1 + \text{tr}(\mathbf{X}^{-1}(\mathbf{A} - \mathbf{X}))}.$$

Derivative of the matrix inverse

Suppose that the invertible matrix \mathbf{A} depends on a parameter t . Then the derivative of the inverse of \mathbf{A} with respect to t is given by

$$\frac{d\mathbf{A}^{-1}}{dt} = -\mathbf{A}^{-1} \frac{d\mathbf{A}}{dt} \mathbf{A}^{-1}.$$

To derive the above expression for the derivative of the inverse of \mathbf{A} , one can differentiate the definition of the matrix inverse $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ and then solve for the inverse of \mathbf{A} :

$$\frac{d\mathbf{A}^{-1}\mathbf{A}}{dt} = \frac{d\mathbf{A}^{-1}}{dt}\mathbf{A} + \mathbf{A}^{-1}\frac{d\mathbf{A}}{dt} = \frac{d\mathbf{I}}{dt} = \mathbf{0}.$$

Subtracting $\mathbf{A}^{-1} \frac{d\mathbf{A}}{dt}$ from both sides of the above and multiplying on the right by \mathbf{A}^{-1} gives the correct expression for the derivative of the inverse:

$$\frac{d\mathbf{A}^{-1}}{dt} = -\mathbf{A}^{-1} \frac{d\mathbf{A}}{dt} \mathbf{A}^{-1}.$$

Similarly, if ϵ is a small number then

$$(\mathbf{A} + \epsilon\mathbf{X})^{-1} = \mathbf{A}^{-1} - \epsilon\mathbf{A}^{-1}\mathbf{X}\mathbf{A}^{-1} + \mathcal{O}(\epsilon^2).$$

Moore–Penrose pseudoinverse

Some of the properties of inverse matrices are shared by Moore–Penrose pseudoinverses, which can be defined for any m -by- n matrix.

Applications

For most practical applications, it is *not* necessary to invert a matrix to solve a system of linear equations; however, for a unique solution, it *is* necessary that the matrix involved be invertible.

Decomposition techniques like LU decomposition are much faster than inversion, and various fast algorithms for special classes of linear systems have also been developed.

Matrix inverses in real-time simulations

Matrix inversion plays a significant role in computer graphics, particularly in 3D graphics rendering and 3D simulations. Examples include screen-to-world ray casting, world-to-subspace-to-world object transformations, and physical simulations.

Matrix inverses in MIMO wireless communication

Matrix inversion also play a significant role in the MIMO (Multiple-Input, Multiple-Output) technology in wireless communications. The MIMO system consists of N transmit and M receive antennas. Unique signals, occupying the same frequency band, are sent via N transmit antennas and are received via M receive antennas. The signal arriving at each receive antenna will be a linear combination of the N transmitted signals forming a $N \times M$ transmission matrix \mathbf{H} . It is crucial for the matrix \mathbf{H} to be invertible for the receiver to be able to figure out the transmitted information.

Notes

[1] , Chapter 2, page 71 (<http://books.google.com/books?id=Gv4pCVyoUVYC&pg=PA71>)

[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *Introduction to Algorithms*, 3rd ed., MIT Press, Cambridge, MA, 2009, §28.2.

[3] Ran Raz. On the complexity of matrix product. In Proceedings of the thirty-fourth annual ACM symposium on Theory of computing. ACM Press, 2002. .

References

- Cormen, Thomas H.; Leiserson, Charles E., Rivest, Ronald L., Stein, Clifford (2001) [1990]. "28.4: Inverting matrices". *Introduction to Algorithms* (2nd ed.). MIT Press and McGraw-Hill. pp. pp. 755–760. ISBN 0-262-03293-7.

External links

- Hazewinkel, Michiel, ed. (2001), "Inversion of a matrix" (<http://www.encyclopediaofmath.org/index.php?title=p/i052440>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Matrix Mathematics: Theory, Facts, and Formulas (<http://books.google.se/books?id=jgEiuHITCYcC&printsec=frontcover>) at Google books
- Equations Solver Online (<http://www.solvingequations.net>)
- Lecture on Inverse Matrices by Khan Academy (<http://www.khanacademy.org/video/inverse-matrix--part-1?playlist=Linear+Algebra>)
- Linear Algebra Lecture on Inverse Matrices by MIT (<http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/lecture-3-multiplication-and-inverse-matrices/>)
- LAPACK (<http://netlib.org/lapack/>) is a collection of FORTRAN subroutines for solving dense linear algebra problems
- ALGLIB (<http://www.alglib.net/eigen/>) includes a partial port of the LAPACK to C++, C#, Delphi, etc.
- Online Inverse Matrix Calculator using AJAX (<http://www.jimmysie.com/math/matrixinv.php>)
- Symbolic Inverse of Matrix Calculator with steps shown (<http://www.emathhelp.net/calculators/linear-algebra/inverse-of-matrix-calculator/>)
- Moore Penrose Pseudoinverse (http://www.vias.org/tmdatanaleng/cc_matrix_pseudoinv.html)
- Inverse of a Matrix Notes (http://numericalmethods.eng.usf.edu/mws/gen/04sle/mws_gen_sle_bck_system.pdf)
- Module for the Matrix Inverse (<http://math.fullerton.edu/mathews/n2003/InverseMatrixMod.html>)
- Calculator for Singular or Non-Square Matrix Inverse (<http://mjollnir.com/matrix/demo.html>)
- Derivative of inverse matrix (<http://planetmath.org/?op=getobj&from=objects&id=6362>), PlanetMath.org.

Eigenvalues and eigenvectors

An **eigenvector** of a square matrix A is a non-zero vector v that, when the matrix is multiplied by v , yields a constant multiple of v , the multiplier being commonly denoted by λ . That is:

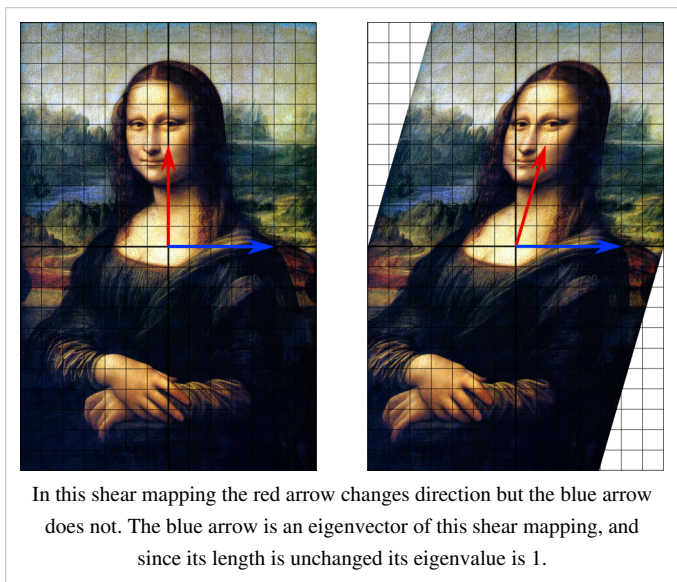
$$Av = \lambda v$$

(Because this equation uses post-multiplication by v , it describes a right eigenvector.)

The number λ is called the **eigenvalue** of A corresponding to v .^[1]

In analytic geometry, for example, a three-element vector may be seen as an arrow in three-dimensional space starting at the origin. In that case, an eigenvector v is an arrow whose direction is either preserved or exactly reversed

after multiplication by A . The corresponding eigenvalue determines how the length of the arrow is changed by the operation, and whether its direction is reversed or not, determined by whether the eigenvalue is negative or positive.



In abstract linear algebra, these concepts are naturally extended to more general situations, where the set of real scalar factors is replaced by any field of scalars (such as algebraic or complex numbers); the set of Cartesian vectors \mathbb{R}^n is replaced by any vector space (such as the continuous functions, the polynomials or the trigonometric series), and matrix multiplication is replaced by any linear operator that maps vectors to vectors (such as the derivative from calculus). In such cases, the "vector" in "eigenvector" may be replaced by a more specific term, such as "eigenfunction", "eigenmode", "eigenface", or "eigenstate". Thus, for example, the exponential function $f(x) = a^x$ is an eigenfunction of the derivative operator " \prime ", with eigenvalue $\lambda = \ln a$, since its derivative is $f'(x) = (\ln a)a^x = \lambda f(x)$.

The set of all eigenvectors of a matrix (or linear operator), each paired with its corresponding eigenvalue, is called the **eigensystem** of that matrix.^[2] Any multiple of an eigenvector is also an eigenvector, with the same eigenvalue. An **eigenspace** of a matrix A is the set of all eigenvectors with the same eigenvalue, together with the zero vector. An **eigenbasis** for A is any basis for the set of all vectors that consists of linearly independent eigenvectors of A . Not every matrix has an eigenbasis, but every symmetric matrix does.

The terms **characteristic vector**, **characteristic value**, and **characteristic space** are also used for these concepts. The prefix **eigen-** is adopted from the German word *eigen* for "self-" or "unique to", "peculiar to", or "belonging to."

Eigenvalues and eigenvectors have many applications in both pure and applied mathematics. They are used in matrix factorization, in quantum mechanics, and in many other areas.

Definition

Eigenvectors and eigenvalues of a real matrix

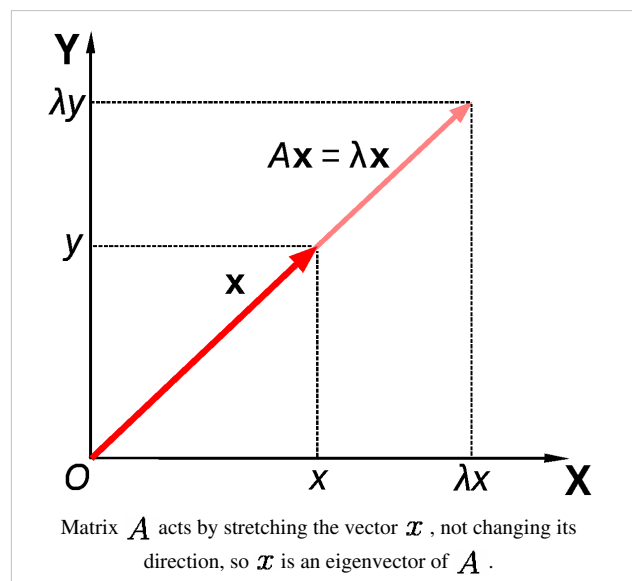
In many contexts, a vector can be assumed to be a list of real numbers (called *elements*), written vertically with brackets around the entire list, such as the vectors u and v below. Two vectors are said to be scalar multiples of each other (also called parallel or collinear) if they have the same number of elements, and if every element of one vector is obtained by multiplying each corresponding element in the other vector by the same number (known as a *scaling factor*, or a *scalar*). For example, the vectors

$$u = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} \text{ and } v = \begin{bmatrix} -20 \\ -60 \\ -80 \end{bmatrix}$$

are scalar multiples of each other, because each element of v is -20 times the corresponding element of u .

A vector with three elements, like u or v above, may represent a point in three-dimensional space, relative to some Cartesian coordinate system. It helps to think of such a vector as the tip of an arrow whose tail is at the origin of the coordinate system. In this case, the condition " u is parallel to v " means that the two arrows lie on the same straight line, and may differ only in length and direction along that line.

If we multiply any square matrix A with n rows and n columns by such a vector v , the result will be another vector $w = Av$, also with n rows and one column. That is,



$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \text{ is mapped to } \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,n} \\ A_{2,1} & A_{2,2} & \dots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \dots & A_{n,n} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

where, for each index i ,

$$w_i = A_{i,1}v_1 + A_{i,2}v_2 + \dots + A_{i,n}v_n = \sum_{j=1}^n A_{i,j}v_j$$

In general, if v_j are not all zeros, the vectors v and Av will not be parallel. When they *are* parallel (that is, when there is some real number λ such that $Av = \lambda v$) we say that v is an **eigenvector** of A . In that case, the scale factor λ is said to be the **eigenvalue** corresponding to that eigenvector.

In particular, multiplication by a 3×3 matrix A may change both the direction and the magnitude of an arrow v in three-dimensional space. However, if v is an eigenvector of A with eigenvalue λ , the operation may only change its length, and either keep its direction or flip it (make the arrow point in the exact opposite direction). Specifically, the length of the arrow will increase if $|\lambda| > 1$, remain the same if $|\lambda| = 1$, and decrease it if $|\lambda| < 1$. Moreover, the direction will be precisely the same if $\lambda > 0$, and flipped if $\lambda < 0$. If $\lambda = 0$, then the length of the arrow becomes zero.

An example

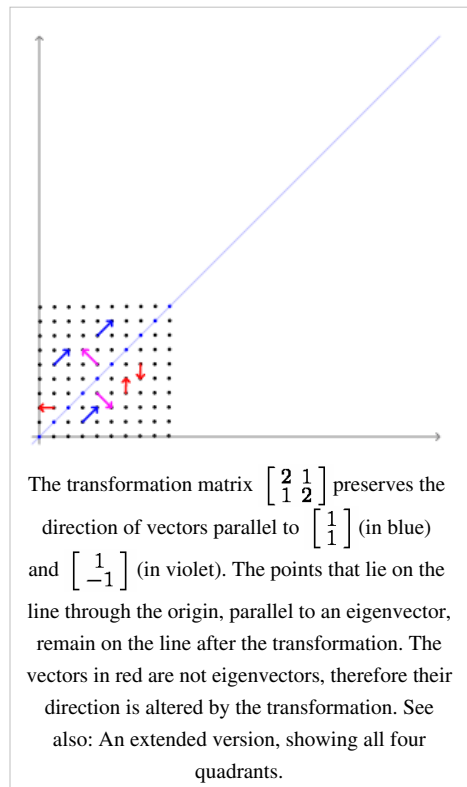
For the transformation matrix

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix},$$

the vector

$$v = \begin{bmatrix} 4 \\ -4 \end{bmatrix}$$

is an eigenvector with eigenvalue 2. Indeed,



$$\begin{aligned} Av &= \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ -4 \end{bmatrix} = \begin{bmatrix} 3 \cdot 4 + 1 \cdot (-4) \\ 1 \cdot 4 + 3 \cdot (-4) \end{bmatrix} \\ &= \begin{bmatrix} 8 \\ -8 \end{bmatrix} = 2 \cdot \begin{bmatrix} 4 \\ -4 \end{bmatrix}. \end{aligned}$$

On the other hand the vector

$$v = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

is *not* an eigenvector, since

$$\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \cdot 0 + 1 \cdot 1 \\ 1 \cdot 0 + 3 \cdot 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix},$$

and this vector is not a multiple of the original vector v .

Another example

For the matrix

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix},$$

we have

$$A \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 1 \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

$$A \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix} = 3 \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Therefore, the vectors $[1, 0, 0]^T$ and $[0, 0, 1]^T$ are eigenvectors of A corresponding to the eigenvalues 1 and 3 respectively. (Here the symbol \top indicates matrix transposition, in this case turning the row vectors into column vectors.)

Trivial cases

The identity matrix I (whose general element I_{ij} is 1 if $i = j$, and 0 otherwise) maps every vector to itself. Therefore, every vector is an eigenvector of I , with eigenvalue 1.

More generally, if A is a diagonal matrix (with $A_{ij} = 0$ whenever $i \neq j$), and v is a vector parallel to axis i (that is, $v_i \neq 0$, and $v_j = 0$ if $j \neq i$), then $Av = \lambda v$ where $\lambda = A_{ii}$. That is, the eigenvalues of a diagonal matrix are the elements of its main diagonal. This is trivially the case of *any* 1×1 matrix.

General definition

The concept of eigenvectors and eigenvalues extends naturally to abstract linear transformations on abstract vector spaces. Namely, let V be any vector space over some field K of scalars, and let T be a linear transformation mapping V into V . We say that a non-zero vector v of V is an **eigenvector** of T if (and only if) there is a scalar λ in K such that

$$T(v) = \lambda v.$$

This equation is called the eigenvalue equation for T , and the scalar λ is the **eigenvalue** of T corresponding to the eigenvector v . Note that $T(v)$ means the result of applying the operator T to the vector v , while λv means the product of the scalar λ by v .^[3]

The matrix-specific definition is a special case of this abstract definition. Namely, the vector space V is the set of all column vectors of a certain size $n \times 1$, and T is the linear transformation that consists in multiplying a vector by the given $n \times n$ matrix A .

Some authors allow v to be the zero vector in the definition of eigenvector. This is reasonable as long as we define eigenvalues and eigenvectors carefully: If we would like the zero vector to be an eigenvector, then we must first

define an eigenvalue of T as a scalar λ in K such that there is a *nonzero* vector v in V with $T(v) = \lambda v$. We then define an eigenvalue of T as a scalar λ in K such that there is a *nonzero* vector v in V with $T(v) = \lambda v$. This way, we ensure that it is not the case that every scalar is an eigenvalue corresponding to the zero vector.

Eigenspace and spectrum

If v is an eigenvector of T , with eigenvalue λ , then any scalar multiple αv of v with nonzero α is also an eigenvector with eigenvalue λ , since $T(\alpha v) = \alpha T(v) = \alpha(\lambda v) = \lambda(\alpha v)$. Moreover, if u and v are eigenvectors with the same eigenvalue λ , then $u + v$ is also an eigenvector with the same eigenvalue λ . Therefore, the set of all eigenvectors with the same eigenvalue λ , together with the zero vector, is a linear subspace of V , called the **eigenspace** of T associated to λ .^[4] If that subspace has dimension 1, it is sometimes called an **eigenline**.^[5]

The **geometric multiplicity** $\gamma_T(\lambda)$ of an eigenvalue λ is the dimension of the eigenspace associated to λ , i.e. number of linearly independent eigenvectors with that eigenvalue.

The eigenspaces of T always form a direct sum (and as a consequence any family of eigenvectors for different eigenvalues is always linearly independent). Therefore the sum of the dimensions of the eigenspaces cannot exceed the dimension n of the space on which T operates, and in particular there cannot be more than n distinct eigenvalues.^[6]

Any subspace spanned by eigenvectors of T is an invariant subspace of T , and the restriction of T to such a subspace is diagonalizable.

The set of eigenvalues of T is sometimes called the spectrum of T .

Eigenbasis

An **eigenbasis** for a linear operator T that operates on a vector space V is a basis for V that consists entirely of eigenvectors of T (possibly with different eigenvalues). Such a basis exists precisely if the direct sum of the eigenspaces equals the whole space, in which case one can take the union of bases chosen in each of the eigenspaces as eigenbasis. The matrix of T in a given basis is diagonal precisely when that basis is an eigenbasis for T , and for this reason T is called **diagonalizable** if it admits an eigenbasis.

Generalizations to infinite-dimensional spaces

The definition of eigenvalue of a linear transformation T remains valid even if the underlying space V is an infinite dimensional Hilbert or Banach space. Namely, a scalar λ is an eigenvalue if and only if there is some nonzero vector v such that $T(v) = \lambda v$.

Eigenfunctions

A widely used class of linear operators acting on infinite dimensional spaces are the differential operators on function spaces. Let D be a linear differential operator in on the space \mathbf{C}^∞ of infinitely differentiable real functions of a real argument t . The eigenvalue equation for D is the differential equation

$$Df = \lambda f$$

The functions that satisfy this equation are commonly called **eigenfunctions**. For the derivative operator d/dt , an eigenfunction is a function that, when differentiated, yields a constant times the original function. If λ is zero, the generic solution is a constant function. If λ is non-zero, the solution is an exponential function

$$f(t) = Ae^{\lambda t}.$$

Eigenfunctions are an essential tool in the solution of differential equations and many other applied and theoretical fields. For instance, the exponential functions are eigenfunctions of any shift invariant linear operator. This fact is

the basis of powerful Fourier transform methods for solving all sorts of problems.

Spectral theory

If λ is an eigenvalue of T , then the operator $T - \lambda I$ is not one-to-one, and therefore its inverse $(T - \lambda I)^{-1}$ is not defined. The converse is true for finite-dimensional vector spaces, but not for infinite-dimensional ones. In general, the operator $T - \lambda I$ may not have an inverse, even if λ is not an eigenvalue.

For this reason, in functional analysis one defines the spectrum of a linear operator T as the set of all scalars λ for which the operator $T - \lambda I$ has no bounded inverse. Thus the spectrum of an operator always contains all its eigenvalues, but is not limited to them.

Associative algebras and representation theory

More algebraically, rather than generalizing the vector space to an infinite dimensional space, one can generalize the algebraic object that is acting on the space, replacing a single operator acting on a vector space with an algebra representation – an associative algebra acting on a module. The study of such actions is the field of representation theory.

A closer analog of eigenvalues is given by the representation-theoretical concept of weight, with the analogs of eigenvectors and eigenspaces being *weight vectors* and *weight spaces*.

Eigenvalues and eigenvectors of matrices

Characteristic polynomial

The eigenvalue equation for a matrix A is

$$Av - \lambda v = 0,$$

which is equivalent to

$$(A - \lambda I)v = 0,$$

where I is the $n \times n$ identity matrix. It is a fundamental result of linear algebra that an equation $Mv = 0$ has a non-zero solution v if, and only if, the determinant $\det(M)$ of the matrix M is zero. It follows that the eigenvalues of A are precisely the real numbers λ that satisfy the equation

$$\det(A - \lambda I) = 0$$

The left-hand side of this equation can be seen (using Leibniz' rule for the determinant) to be a polynomial function of the variable λ . The degree of this polynomial is n , the order of the matrix. Its coefficients depend on the entries of A , except that its term of degree n is always $(-1)^n \lambda^n$. This polynomial is called the *characteristic polynomial* of A ; and the above equation is called the *characteristic equation* (or, less often, the *secular equation*) of A .

For example, let A be the matrix

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix}$$

The characteristic polynomial of A is

$$\det(A - \lambda I) = \det \left(\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) = \det \begin{bmatrix} 2 - \lambda & 0 & 0 \\ 0 & 3 - \lambda & 4 \\ 0 & 4 & 9 - \lambda \end{bmatrix}$$

which is

$$(2 - \lambda)[(3 - \lambda)(9 - \lambda) - 16] = -\lambda^3 + 14\lambda^2 - 35\lambda + 22$$

The roots of this polynomial are 2, 1, and 11. Indeed these are the only three eigenvalues of A , corresponding to the eigenvectors $[1, 0, 0]'$, $[0, 2, -1]'$, and $[0, 1, 2]'$ (or any non-zero multiples thereof).

In the real domain

Since the eigenvalues are roots of the characteristic polynomial, an $n \times n$ matrix has at most n eigenvalues. If the matrix has real entries, the coefficients of the characteristic polynomial are all real; but it may have fewer than n real roots, or no real roots at all.

For example, consider the cyclic permutation matrix

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

This matrix shifts the coordinates of the vector up by one position, and moves the first coordinate to the bottom. Its characteristic polynomial is $1 - \lambda^3$ which has one real root $\lambda_1 = 1$. Any vector with three equal non-zero elements is an eigenvector for this eigenvalue. For example,

$$A \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} = 1 \cdot \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix}$$

In the complex domain

The fundamental theorem of algebra implies that the characteristic polynomial of an $n \times n$ matrix A , being a polynomial of degree n , has exactly n complex roots. More precisely, it can be factored into the product of n linear terms,

$$\det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda)$$

where each λ_i is a complex number. The numbers $\lambda_1, \lambda_2, \dots, \lambda_n$, (which may not be all distinct) are roots of the polynomial, and are precisely the eigenvalues of A .

Even if the entries of A are all real numbers, the eigenvalues may still have non-zero imaginary parts (and the elements of the corresponding eigenvectors will therefore also have non-zero imaginary parts). Also, the eigenvalues may be irrational numbers even if all the entries of A are rational numbers, or all are integers. However, if the entries of A are algebraic numbers (which include the rationals), the eigenvalues will be (complex) algebraic numbers too.

The non-real roots of a real polynomial with real coefficients can be grouped into pairs of complex conjugate values, namely with the two members of each pair having the same real part and imaginary parts that differ only in sign. If the degree is odd, then by the intermediate value theorem at least one of the roots will be real. Therefore, any real matrix with odd order will have at least one real eigenvalue; whereas a real matrix with even order may have no real eigenvalues.

In the example of the 3×3 cyclic permutation matrix A , above, the characteristic polynomial $1 - \lambda^3$ has two additional non-real roots, namely

$$\lambda_2 = -1/2 + i\sqrt{3}/2 \text{ and } \lambda_3 = \lambda_2^* = -1/2 - i\sqrt{3}/2,$$

where $i = \sqrt{-1}$ is the imaginary unit. Note that $\lambda_2\lambda_3 = 1$, $\lambda_2^2 = \lambda_3$, and $\lambda_3^2 = \lambda_2$. Then

$$A \begin{bmatrix} 1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} \lambda_2 \\ \lambda_3 \\ 1 \end{bmatrix} = \lambda_2 \cdot \begin{bmatrix} 1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \text{ and } A \begin{bmatrix} 1 \\ \lambda_3 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \lambda_3 \\ \lambda_2 \\ 1 \end{bmatrix} = \lambda_3 \cdot \begin{bmatrix} 1 \\ \lambda_3 \\ \lambda_2 \end{bmatrix}$$

Therefore, the vectors $[1, \lambda_2, \lambda_3]'$ and $[1, \lambda_3, \lambda_2]'$ are eigenvectors of A , with eigenvalues λ_2 , and λ_3 , respectively.

Algebraic multiplicities

Let λ_i be an eigenvalue of an $n \times n$ matrix A . The *algebraic multiplicity* $\mu_A(\lambda_i)$ of λ_i is its multiplicity as a root of the characteristic polynomial, that is, the largest integer k such that $(\lambda - \lambda_i)^k$ divides evenly that polynomial.

Like the geometric multiplicity $\gamma_A(\lambda_i)$, the algebraic multiplicity is an integer between 1 and n ; and the sum μ_A of $\mu_A(\lambda_i)$ over all *distinct* eigenvalues also cannot exceed n . If complex eigenvalues are considered, μ_A is exactly n .

It can be proved that the geometric multiplicity $\gamma_A(\lambda_i)$ of an eigenvalue never exceeds its algebraic multiplicity $\mu_A(\lambda_i)$. Therefore, γ_A is at most μ_A .

If $\gamma_A(\lambda_i) = \mu_A(\lambda_i)$, then λ_i is said to be a *semisimple eigenvalue*.

Example

For the matrix: $A = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 3 \end{bmatrix}$,

the characteristic polynomial of A is

$$\det(A - \lambda I) = \det \begin{bmatrix} 2 - \lambda & 0 & 0 & 0 \\ 1 & 2 - \lambda & 0 & 0 \\ 0 & 1 & 3 - \lambda & 0 \\ 0 & 0 & 1 & 3 - \lambda \end{bmatrix} = (2 - \lambda)^2(3 - \lambda)^2,$$

being the product of the diagonal with a lower triangular matrix.

The roots of this polynomial, and hence the eigenvalues, are 2 and 3. The *algebraic multiplicity* of each eigenvalue is 2; in other words they are both double roots. On the other hand, the *geometric multiplicity* of the eigenvalue 2 is only 1, because its eigenspace is spanned by the vector $[0, 1, -1, 1]$, and is therefore 1 dimensional. Similarly, the geometric multiplicity of the eigenvalue 3 is 1 because its eigenspace is spanned by $[0, 0, 0, 1]$. Hence, the total algebraic multiplicity of A , denoted μ_A , is 4, which is the most it could be for a 4 by 4 matrix. The geometric multiplicity γ_A is 2, which is the smallest it could be for a matrix which has two distinct eigenvalues.

Diagonalization and eigendecomposition

If the sum γ_A of the geometric multiplicities of all eigenvalues is exactly n , then A has a set of n linearly independent eigenvectors. Let Q be a square matrix whose columns are those eigenvectors, in any order. Then we will have $AQ = Q\Lambda$, where Λ is the diagonal matrix such that Λ_{ii} is the eigenvalue associated to column i of Q . Since the columns of Q are linearly independent, the matrix Q is invertible. Premultiplying both sides by Q^{-1} we get $Q^{-1}AQ = \Lambda$. By definition, therefore, the matrix A is diagonalizable.

Conversely, if A is diagonalizable, let Q be a non-singular square matrix such that $Q^{-1}AQ$ is some diagonal matrix D . Multiplying both sides on the left by Q we get $AQ = QD$. Therefore each column of Q must be an eigenvector of A , whose eigenvalue is the corresponding element on the diagonal of D . Since the columns of Q must be linearly independent, it follows that $\gamma_A = n$. Thus γ_A is equal to n if and only if A is diagonalizable.

If A is diagonalizable, the space of all n -element vectors can be decomposed into the direct sum of the eigenspaces of A . This decomposition is called the eigendecomposition of A , and it is preserved under change of

coordinates.

A matrix that is not diagonalizable is said to be defective. For defective matrices, the notion of eigenvector can be generalized to generalized eigenvectors, and that of diagonal matrix to a Jordan form matrix. Over an algebraically closed field, any matrix A has a Jordan form and therefore admits a basis of generalized eigenvectors, and a decomposition into generalized eigenspaces

Further properties

Let A be an arbitrary $n \times n$ matrix of complex numbers with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$. (Here it is understood that an eigenvalue with algebraic multiplicity μ occurs μ times in this list.) Then

- The trace of A , defined as the sum of its diagonal elements, is also the sum of all eigenvalues:

$$\operatorname{tr}(A) = \sum_{i=1}^n A_{ii} = \sum_{i=1}^n \lambda_i = \lambda_1 + \lambda_2 + \dots + \lambda_n.$$

- The determinant of A is the product of all eigenvalues:

$$\det(A) = \prod_{i=1}^n \lambda_i = \lambda_1 \lambda_2 \dots \lambda_n.$$

- The eigenvalues of the k th power of A , i.e. the eigenvalues of A^k , for any positive integer k , are $\lambda_1^k, \lambda_2^k, \dots, \lambda_n^k$.
- The matrix A is invertible if and only if all the eigenvalues λ_i are nonzero.
- If A is invertible, then the eigenvalues of A^{-1} are $1/\lambda_1, 1/\lambda_2, \dots, 1/\lambda_n$.
- If A is equal to its conjugate transpose A^* (in other words, if A is Hermitian), then every eigenvalue is real. The same is true of any a symmetric real matrix. If A is also positive-definite, positive-semidefinite, negative-definite, or negative-semidefinite every eigenvalue is positive, non-negative, negative, or non-positive respectively.
- Every eigenvalue of a unitary matrix has absolute value $|\lambda| = 1$.

Left and right eigenvectors

The use of matrices with a single column (rather than a single row) to represent vectors is traditional in many disciplines. For that reason, the word "eigenvector" almost always means a **right eigenvector**, namely a *column* vector that must be placed to the *right* of the matrix A in the defining equation

$$Av = \lambda v.$$

There may be also single-*row* vectors that are unchanged when they occur on the *left* side of a product with a square matrix A ; that is, which satisfy the equation

$$uA = \lambda u$$

Any such row vector u is called a **left eigenvector** of A .

The left eigenvectors of A are transposes of the right eigenvectors of the transposed matrix A^T , since their defining equation is equivalent to

$$A^T u^T = \lambda u^T$$

It follows that, if A is Hermitian, its left and right eigenvectors are complex conjugates. In particular if A is a real symmetric matrix, they are the same except for transposition.

Calculation

Computing the eigenvalues

The eigenvalues of a matrix A can be determined by finding the roots of the characteristic polynomial. Explicit algebraic formulas for the roots of a polynomial exist only if the degree n is 4 or less. According to the Abel–Ruffini theorem there is no general, explicit and exact algebraic formula for the roots of a polynomial with degree 5 or more.

It turns out that any polynomial with degree n is the characteristic polynomial of some companion matrix of order n . Therefore, for matrices of order 5 or more, the eigenvalues and eigenvectors cannot be obtained by an explicit algebraic formula, and must therefore be computed by approximate numerical methods.

In theory, the coefficients of the characteristic polynomial can be computed exactly, since they are sums of products of matrix elements; and there are algorithms that can find all the roots of a polynomial of arbitrary degree to any required accuracy. However, this approach is not viable in practice because the coefficients would be contaminated by unavoidable round-off errors, and the roots of a polynomial can be an extremely sensitive function of the coefficients (as exemplified by Wilkinson's polynomial).

Efficient, accurate methods to compute eigenvalues and eigenvectors of arbitrary matrices were not known until the advent of the QR algorithm in 1961. Combining the Householder transformation with the LU decomposition results in an algorithm with better convergence than the QR algorithm.^[citation needed] For large Hermitian sparse matrices, the Lanczos algorithm is one example of an efficient iterative method to compute eigenvalues and eigenvectors, among several other possibilities.

Computing the eigenvectors

Once the (exact) value of an eigenvalue is known, the corresponding eigenvectors can be found by finding non-zero solutions of the eigenvalue equation, that becomes a system of linear equations with known coefficients. For example, once it is known that 6 is an eigenvalue of the matrix

$$A = \begin{bmatrix} 4 & 1 \\ 6 & 3 \end{bmatrix}$$

we can find its eigenvectors by solving the equation $Av = 6v$, that is

$$\begin{bmatrix} 4 & 1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 6 \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

This matrix equation is equivalent to two linear equations

$$\begin{cases} 4x + y = 6x \\ 6x + 3y = 6y \end{cases} \text{ that is } \begin{cases} -2x + y = 0 \\ +6x - 3y = 0 \end{cases}$$

Both equations reduce to the single linear equation $y = 2x$. Therefore, any vector of the form $[a, 2a]'$, for any non-zero real number a , is an eigenvector of A with eigenvalue $\lambda = 6$.

The matrix A above has another eigenvalue $\lambda = 1$. A similar calculation shows that the corresponding eigenvectors are the non-zero solutions of $3x + y = 0$, that is, any vector of the form $[b, -3b]'$, for any non-zero real number b .

Some numeric methods that compute the eigenvalues of a matrix also determine a set of corresponding eigenvectors as a by-product of the computation.

History

Eigenvalues are often introduced in the context of linear algebra or matrix theory. Historically, however, they arose in the study of quadratic forms and differential equations.

In the 18th century Euler studied the rotational motion of a rigid body and discovered the importance of the principal axes. Lagrange realized that the principal axes are the eigenvectors of the inertia matrix.^[7] In the early 19th century, Cauchy saw how their work could be used to classify the quadric surfaces, and generalized it to arbitrary dimensions.^[8] Cauchy also coined the term *racine caractéristique* (characteristic root) for what is now called *eigenvalue*; his term survives in *characteristic equation*.^[9]

Fourier used the work of Laplace and Lagrange to solve the heat equation by separation of variables in his famous 1822 book *Théorie analytique de la chaleur*.^[10] Sturm developed Fourier's ideas further and brought them to the attention of Cauchy, who combined them with his own ideas and arrived at the fact that real symmetric matrices have real eigenvalues. This was extended by Hermite in 1855 to what are now called Hermitian matrices. Around the same time, Brioschi proved that the eigenvalues of orthogonal matrices lie on the unit circle, and Clebsch found the corresponding result for skew-symmetric matrices. Finally, Weierstrass clarified an important aspect in the stability theory started by Laplace by realizing that defective matrices can cause instability.

In the meantime, Liouville studied eigenvalue problems similar to those of Sturm; the discipline that grew out of their work is now called *Sturm–Liouville theory*.^[11] Schwarz studied the first eigenvalue of Laplace's equation on general domains towards the end of the 19th century, while Poincaré studied Poisson's equation a few years later.^[12]

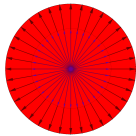
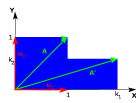
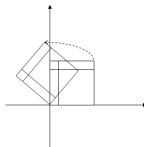
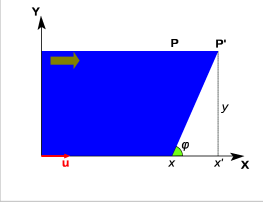
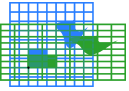
At the start of the 20th century, Hilbert studied the eigenvalues of integral operators by viewing the operators as infinite matrices.^[13] He was the first to use the German word *eigen* to denote eigenvalues and eigenvectors in 1904, though he may have been following a related usage by Helmholtz. For some time, the standard term in English was "proper value", but the more distinctive term "eigenvalue" is standard today.^[14]

The first numerical algorithm for computing eigenvalues and eigenvectors appeared in 1929, when Von Mises published the power method. One of the most popular methods today, the QR algorithm, was proposed independently by John G.F. Francis^[15] and Vera Kublanovskaya^[16] in 1961.^[17]

Applications

Eigenvalues of geometric transformations

The following table presents some example transformations in the plane along with their 2x2 matrices, eigenvalues, and eigenvectors.

	scaling	unequal scaling	rotation	horizontal shear	hyperbolic rotation
illustration					
matrix	$\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}$	$\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$	$\begin{bmatrix} c & -s \\ s & c \end{bmatrix}$ $c = \cos \theta$ $s = \sin \theta$	$\begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} c & s \\ s & c \end{bmatrix}$ $c = \cosh \varphi$ $s = \sinh \varphi$
characteristic polynomial	$(\lambda - k)^2$	$(\lambda - k_1)(\lambda - k_2)$	$\lambda^2 - 2c\lambda + 1$	$(\lambda - 1)^2$	$\lambda^2 - 2c\lambda + 1$

eigenvalues λ_i	$\lambda_1 = \lambda_2 = k$	$\lambda_1 = k_1$ $\lambda_2 = k_2$	$\lambda_1 = e^{i\theta} = c + s\mathbf{i}$ $\lambda_2 = e^{-i\theta} = c - s\mathbf{i}$	$\lambda_1 = \lambda_2 = 1$	$\lambda_1 = e^\varphi$ $\lambda_2 = e^{-\varphi}$
algebraic multipl. $\mu_i = \mu(\lambda_i)$	$\mu_1 = 2$	$\mu_1 = 1$ $\mu_2 = 1$	$\mu_1 = 1$ $\mu_2 = 1$	$\mu_1 = 2$	$\mu_1 = 1$ $\mu_2 = 1$
geometric multipl. $\gamma_i = \gamma(\lambda_i)$	$\gamma_1 = 2$	$\gamma_1 = 1$ $\gamma_2 = 1$	$\gamma_1 = 1$ $\gamma_2 = 1$	$\gamma_1 = 1$	$\gamma_1 = 1$ $\gamma_2 = 1$
eigenvectors	All non-zero vectors	$u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $u_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$u_1 = \begin{bmatrix} 1 \\ -\mathbf{i} \end{bmatrix}$ $u_2 = \begin{bmatrix} 1 \\ +\mathbf{i} \end{bmatrix}$	$u_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$u_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $u_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

Note that the characteristic equation for a rotation is a quadratic equation with discriminant $D = -4(\sin \theta)^2$, which is a negative number whenever θ is not an integer multiple of 180° . Therefore, except for these special cases, the two eigenvalues are complex numbers, $\cos \theta \pm \mathbf{i} \sin \theta$; and all eigenvectors have non-real entries. Indeed, except for those special cases, a rotation changes the direction of every nonzero vector in the plane.

Schrödinger equation

An example of an eigenvalue equation where the transformation T is represented in terms of a differential operator is the time-independent Schrödinger equation in quantum mechanics:

$$H\psi_E = E\psi_E$$

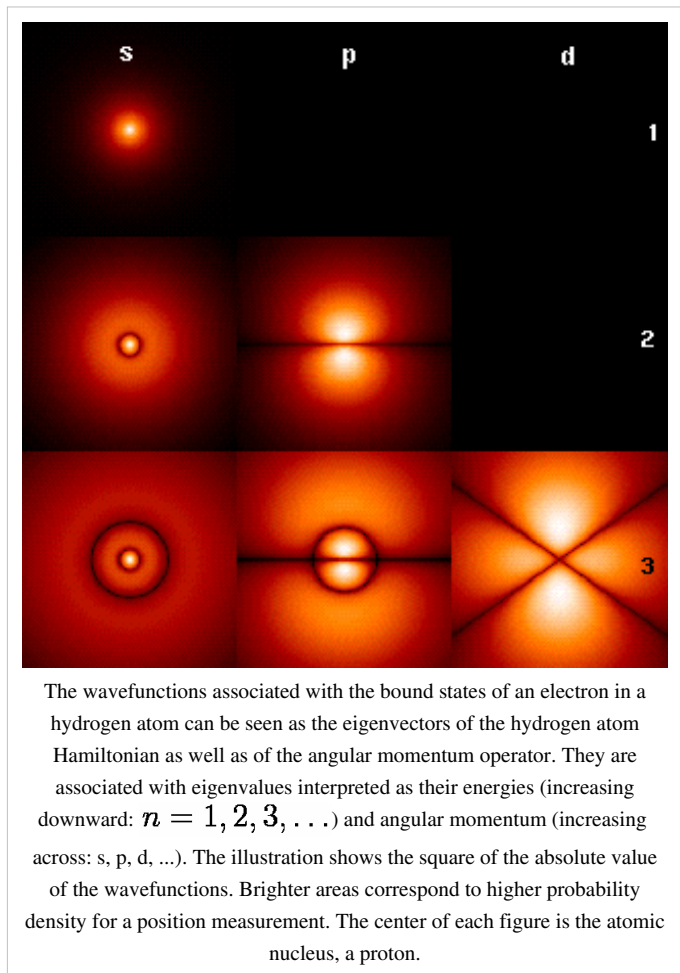
where H , the Hamiltonian, is a second-order differential operator and ψ_E , the wavefunction, is one of its eigenfunctions corresponding to the eigenvalue E , interpreted as its energy.

However, in the case where one is interested only in the bound state solutions of the Schrödinger equation, one looks for ψ_E within the space of square integrable functions. Since this space is a Hilbert space with a well-defined scalar product, one can introduce a basis set in which ψ_E and H can be represented as a one-dimensional array and a matrix respectively. This allows one to represent the Schrödinger equation in a matrix form.

Bra-ket notation is often used in this context. A vector, which represents a state of the system, in the Hilbert space of square integrable functions is represented by $|\Psi_E\rangle$. In this notation, the Schrödinger equation is:

$$H|\Psi_E\rangle = E|\Psi_E\rangle$$

where $|\Psi_E\rangle$ is an **eigenstate** of H . It is a self adjoint operator, the infinite dimensional analog of Hermitian matrices (see *Observable*). As in the matrix case, in the equation above $H|\Psi_E\rangle$ is understood to be the vector obtained by application of the transformation H to $|\Psi_E\rangle$.



Molecular orbitals

In quantum mechanics, and in particular in atomic and molecular physics, within the Hartree–Fock theory, the atomic and molecular orbitals can be defined by the eigenvectors of the Fock operator. The corresponding eigenvalues are interpreted as ionization potentials via Koopmans' theorem. In this case, the term eigenvector is used in a somewhat more general meaning, since the Fock operator is explicitly dependent on the orbitals and their eigenvalues. If one wants to underline this aspect one speaks of nonlinear eigenvalue problem. Such equations are usually solved by an iteration procedure, called in this case self-consistent field method. In quantum chemistry, one often represents the Hartree–Fock equation in a non-orthogonal basis set. This particular representation is a generalized eigenvalue problem called Roothaan equations.

Geology and glaciology

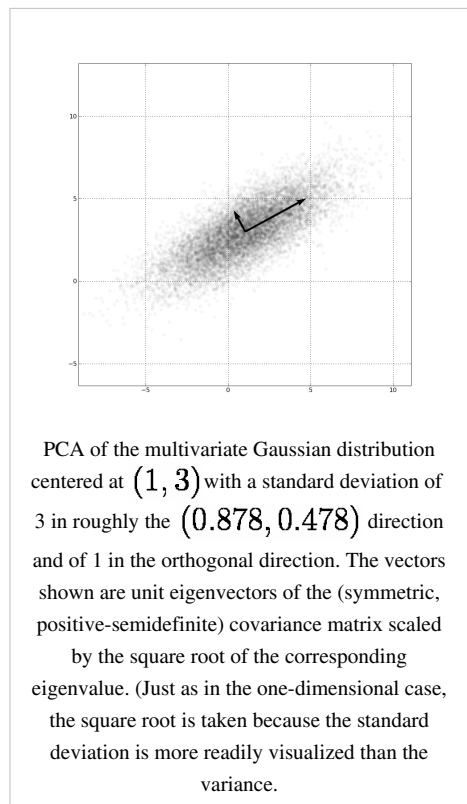
In geology, especially in the study of glacial till, eigenvectors and eigenvalues are used as a method by which a mass of information of a clast fabric's constituents' orientation and dip can be summarized in a 3-D space by six numbers. In the field, a geologist may collect such data for hundreds or thousands of clasts in a soil sample, which can only be compared graphically such as in a Tri-Plot (Sneed and Folk) diagram, or as a Stereonet on a Wulff Net.

The output for the orientation tensor is in the three orthogonal (perpendicular) axes of space. The three eigenvectors are ordered v_1, v_2, v_3 by their eigenvalues $E_1 \geq E_2 \geq E_3$,^[18] v_1 then is the primary orientation/dip of clast, v_2 is the secondary and v_3 is the tertiary, in terms of strength. The clast orientation is defined as the direction of the eigenvector, on a compass rose of 360°. Dip is measured as the eigenvalue, the modulus of the tensor: this is valued from 0° (no dip) to 90° (vertical). The relative values of E_1, E_2 , and E_3 are dictated by the nature of the sediment's fabric. If $E_1 = E_2 = E_3$, the fabric is said to be isotropic. If $E_1 = E_2 > E_3$, the fabric is said to be planar. If $E_1 > E_2 > E_3$, the fabric is said to be linear.

Principal components analysis

The eigendecomposition of a symmetric positive semidefinite (PSD) matrix yields an orthogonal basis of eigenvectors, each of which has a nonnegative eigenvalue. The orthogonal decomposition of a PSD matrix is used in multivariate analysis, where the sample covariance matrices are PSD. This orthogonal decomposition is called principal components analysis (PCA) in statistics. PCA studies linear relations among variables. PCA is performed on the covariance matrix or the correlation matrix (in which each variable is scaled to have its sample variance equal to one). For the covariance or correlation matrix, the eigenvectors correspond to principal components and the eigenvalues to the variance explained by the principal components. Principal component analysis of the correlation matrix provides an orthonormal eigen-basis for the space of the observed data: In this basis, the largest eigenvalues correspond to the principal-components that are associated with most of the covariability among a number of observed data.

Principal component analysis is used to study large data sets, such as those encountered in data mining, chemical research, psychology, and in marketing. PCA is popular especially in psychology, in the field of psychometrics. In Q methodology, the eigenvalues of the correlation



matrix determine the Q-methodologist's judgment of *practical* significance (which differs from the statistical significance of hypothesis testing; cf. criteria for determining the number of factors). More generally, principal component analysis can be used as a method of factor analysis in structural equation modeling.

Vibration analysis

Eigenvalue problems occur naturally in the vibration analysis of mechanical structures with many degrees of freedom. The eigenvalues are used to determine the natural frequencies (or **eigenfrequencies**) of vibration, and the eigenvectors determine the shapes of these vibrational modes. In particular, undamped vibration is governed by

$$m\ddot{x} + kx = 0$$

or

$$m\ddot{x} = -kx$$

that is, acceleration is proportional to position (i.e., we expect x to be sinusoidal in time).

In n dimensions, m becomes a mass matrix and k a stiffness matrix. Admissible solutions are then a linear combination of solutions to the generalized eigenvalue problem

$$-kx = \omega^2 mx$$

where ω^2 is the eigenvalue and ω is the angular frequency. Note that the principal vibration modes are different from the principal compliance modes, which are the eigenvectors of k alone. Furthermore, damped vibration, governed by

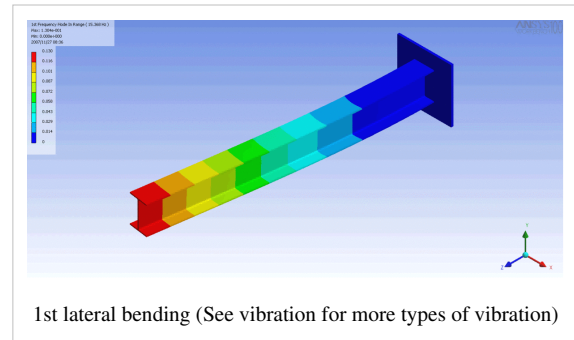
$$m\ddot{x} + c\dot{x} + kx = 0$$

leads to what is called a so-called quadratic eigenvalue problem,

$$(\omega^2 m + \omega c + k)x = 0.$$

This can be reduced to a generalized eigenvalue problem by clever use of algebra at the cost of solving a larger system.

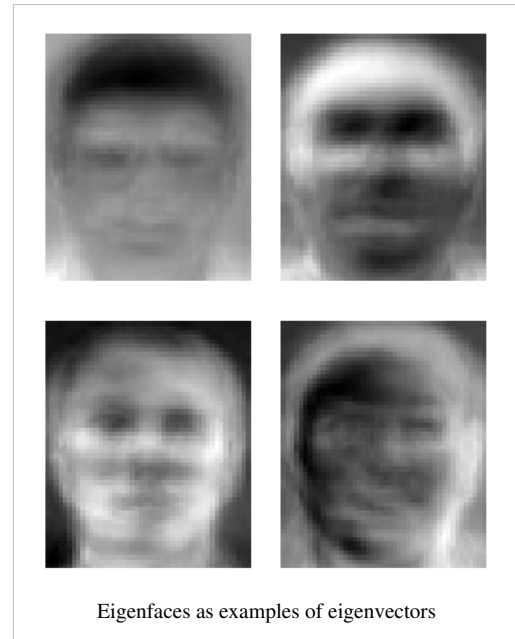
The orthogonality properties of the eigenvectors allows decoupling of the differential equations so that the system can be represented as linear summation of the eigenvectors. The eigenvalue problem of complex structures is often solved using finite element analysis, but neatly generalize the solution to scalar-valued vibration problems.



Eigenfaces

In image processing, processed images of faces can be seen as vectors whose components are the brightnesses of each pixel. The dimension of this vector space is the number of pixels. The eigenvectors of the covariance matrix associated with a large set of normalized pictures of faces are called **eigenfaces**; this is an example of principal components analysis. They are very useful for expressing any face image as a linear combination of some of them. In the facial recognition branch of biometrics, eigenfaces provide a means of applying data compression to faces for identification purposes. Research related to eigen vision systems determining hand gestures has also been made.

Similar to this concept, **eigenvoices** represent the general direction of variability in human pronunciations of a particular utterance, such as a word in a language. Based on a linear combination of such eigenvoices, a new voice pronunciation of the word can be constructed. These concepts have been found useful in automatic speech recognition systems, for speaker adaptation.



Tensor of moment of inertia

In mechanics, the eigenvectors of the moment of inertia tensor define the principal axes of a rigid body. The tensor of moment of inertia is a key quantity required to determine the rotation of a rigid body around its center of mass.

Stress tensor

In solid mechanics, the stress tensor is symmetric and so can be decomposed into a diagonal tensor with the eigenvalues on the diagonal and eigenvectors as a basis. Because it is diagonal, in this orientation, the stress tensor has no shear components; the components it does have are the principal components.

Eigenvalues of a graph

In spectral graph theory, an eigenvalue of a graph is defined as an eigenvalue of the graph's adjacency matrix A , or (increasingly) of the graph's Laplacian matrix (see also Discrete Laplace operator), which is either $T - A$ (sometimes called the *combinatorial Laplacian*) or $I - T^{-1/2}AT^{-1/2}$ (sometimes called the *normalized Laplacian*), where T is a diagonal matrix with T_{ii} equal to the degree of vertex v_i , and in $T^{-1/2}$, the i th diagonal entry is $\sqrt{\deg(v_i)}$. The k th principal eigenvector of a graph is defined as either the eigenvector corresponding to the k th largest or k th smallest eigenvalue of the Laplacian. The first principal eigenvector of the graph is also referred to merely as the principal eigenvector.

The principal eigenvector is used to measure the centrality of its vertices. An example is Google's PageRank algorithm. The principal eigenvector of a modified adjacency matrix of the World Wide Web graph gives the page ranks as its components. This vector corresponds to the stationary distribution of the Markov chain represented by the row-normalized adjacency matrix; however, the adjacency matrix must first be modified to ensure a stationary distribution exists. The second smallest eigenvector can be used to partition the graph into clusters, via spectral clustering. Other methods are also available for clustering.

Basic reproduction number

See Basic reproduction number

The basic reproduction number (R_0) is a fundamental number in the study of how infectious diseases spread. If one infectious person is put into a population of completely susceptible people, then R_0 is the average number of people that one typical infectious person will infect. The generation time of an infection is the time, t_G , from one person becoming infected to the next person becoming infected. In a heterogenous population, the next generation matrix defines how many people in the population will become infected after time t_G has passed. R_0 is then the largest eigenvalue of the next generation matrix.

Notes

- [1] Wolfram Research, Inc. (2010) *Eigenvector* (<http://mathworld.wolfram.com/Eigenvector.html>). Accessed on 2010-01-29.
- [2] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery (2007), *Numerical Recipes: The Art of Scientific Computing*, Chapter 11: *Eigensystems.*, pages=563–597. Third edition, Cambridge University Press. ISBN 9780521880688 (<http://www.nr.com/>)
- [3] See ;
- [4] Lemma for the eigenspace
- [5] *Schaum's Easy Outline of Linear Algebra* (<http://books.google.com/books?id=pkESXAcliCQC&pg=PA111>), p. 111
- [6] For a proof of this lemma, see ; ; ; and Lemma for linear independence of eigenvectors
- [7] See
- [8] See
- [9] See
- [10] See
- [11] See
- [12] See
- [13] See
- [14] See
- [15] and
- [16] . Also published in:
- [17] See ;
- [18] Stereo32 software (<http://www.ruhr-uni-bochum.de/hardrock/downloads.htm>)

References

- Korn, Granino A.; Korn, Theresa M. (2000), "Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review", *New York: McGraw-Hill* (1152 p., Dover Publications, 2 Revised edition), Bibcode: 1968mhse.book.....K (<http://adsabs.harvard.edu/abs/1968mhse.book.....K>), ISBN 0-486-41147-8.
- Lipschutz, Seymour (1991), *Schaum's outline of theory and problems of linear algebra*, Schaum's outline series (2nd ed.), New York, NY: McGraw-Hill Companies, ISBN 0-07-038007-4.
- Friedberg, Stephen H.; Insel, Arnold J.; Spence, Lawrence E. (1989), *Linear algebra* (2nd ed.), Englewood Cliffs, NJ 07632: Prentice Hall, ISBN 0-13-537102-3.
- Aldrich, John (2006), "Eigenvalue, eigenfunction, eigenvector, and related terms" (<http://jeff560.tripod.com/e.html>), in Jeff Miller (Editor), *Earliest Known Uses of Some of the Words of Mathematics* (<http://jeff560.tripod.com/e.html>), retrieved 2006-08-22
- Strang, Gilbert (1993), *Introduction to linear algebra*, Wellesley-Cambridge Press, Wellesley, MA, ISBN 0-9614088-5-5.
- Strang, Gilbert (2006), *Linear algebra and its applications*, Thomson, Brooks/Cole, Belmont, CA, ISBN 0-03-010567-6.
- Bowen, Ray M.; Wang, Chao-Cheng (1980), *Linear and multilinear algebra*, Plenum Press, New York, NY, ISBN 0-306-37508-7.

- Cohen-Tannoudji, Claude (1977), "Chapter II. The mathematical tools of quantum mechanics", *Quantum mechanics*, John Wiley & Sons, ISBN 0-471-16432-1.
- Fraleigh, John B.; Bearegard, Raymond A. (1995), *Linear algebra* (3rd ed.), Addison-Wesley Publishing Company, ISBN 0-201-83999-7 (international edition) Check `| isbn= value (help)`.
- Golub, Gene H.; Van Loan, Charles F. (1996), *Matrix computations (3rd Edition)*, Johns Hopkins University Press, Baltimore, MD, ISBN 978-0-8018-5414-9.
- Hawkins, T. (1975), "Cauchy and the spectral theory of matrices", *Historia Mathematica* **2**: 1–29, doi: 10.1016/0315-0860(75)90032-4 ([http://dx.doi.org/10.1016/0315-0860\(75\)90032-4](http://dx.doi.org/10.1016/0315-0860(75)90032-4)).
- Horn, Roger A.; Johnson, Charles F. (1985), *Matrix analysis*, Cambridge University Press, ISBN 0-521-30586-1 (hardback), ISBN 0-521-38632-2 (paperback) Check `| isbn= value (help)`.
- Kline, Morris (1972), *Mathematical thought from ancient to modern times*, Oxford University Press, ISBN 0-19-501496-0.
- Meyer, Carl D. (2000), *Matrix analysis and applied linear algebra*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, ISBN 978-0-89871-454-8.
- Brown, Maureen (October 2004), *Illuminating Patterns of Perception: An Overview of Q Methodology*.
- Golub, Gene F.; van der Vorst, Henk A. (2000), "Eigenvalue computation in the 20th century", *Journal of Computational and Applied Mathematics* **123**: 35–65, doi: 10.1016/S0377-0427(00)00413-1 ([http://dx.doi.org/10.1016/S0377-0427\(00\)00413-1](http://dx.doi.org/10.1016/S0377-0427(00)00413-1)).
- Akivis, Max A.; Vladislav V. Goldberg (1969), *Tensor calculus*, Russian, Science Publishers, Moscow .
- Gelfand, I. M. (1971), *Lecture notes in linear algebra*, Russian, Science Publishers, Moscow.
- Alexandrov, Pavel S. (1968), *Lecture notes in analytical geometry*, Russian, Science Publishers, Moscow.
- Carter, Tamara A.; Tapia, Richard A.; Papaconstantinou, Anne, *Linear Algebra: An Introduction to Linear Algebra for Pre-Calculus Students* (<http://ceee.rice.edu/Books/LA/index.html>), Rice University, Online Edition, retrieved 2008-02-19.
- Roman, Steven (2008), *Advanced linear algebra* (3rd ed.), New York, NY: Springer Science + Business Media, LLC, ISBN 978-0-387-72828-5.
- Shilov, Georgi E. (1977), *Linear algebra* (translated and edited by Richard A. Silverman ed.), New York: Dover Publications, ISBN 0-486-63518-X.
- Hefferon, Jim (2001), *Linear Algebra* (<http://joshua.smcvt.edu/linearalgebra/>), Online book, St Michael's College, Colchester, Vermont, USA.
- Kuttler, Kenneth (2007), *An introduction to linear algebra* (<http://www.math.byu.edu/~klkuttler/Linearalgebra.pdf>) (PDF), Online e-book in PDF format, Brigham Young University.
- Demmel, James W. (1997), *Applied numerical linear algebra*, SIAM, ISBN 0-89871-389-7.
- Beezer, Robert A. (2006), *A first course in linear algebra* (<http://linear.ups.edu/>), Free online book under GNU licence, University of Puget Sound.
- Lancaster, P. (1973), *Matrix theory*, Russian, Moscow, Russia: Science Publishers.
- Halmos, Paul R. (1987), *Finite-dimensional vector spaces* (8th ed.), New York, NY: Springer-Verlag, ISBN 0-387-90093-4.
- Pigolkina, T. S. and Shulman, V. S., *Eigenvalue* (in Russian), In: Vinogradov, I. M. (Ed.), *Mathematical Encyclopedia*, Vol. 5, Soviet Encyclopedia, Moscow, 1977.
- Greub, Werner H. (1975), *Linear Algebra (4th Edition)*, Springer-Verlag, New York, NY, ISBN 0-387-90110-8.
- Larson, Ron; Edwards, Bruce H. (2003), *Elementary linear algebra* (5th ed.), Houghton Mifflin Company, ISBN 0-618-33567-6.
- Curtis, Charles W., *Linear Algebra: An Introductory Approach*, 347 p., Springer; 4th ed. 1984. Corr. 7th printing edition (August 19, 1999), ISBN 0-387-90992-3.
- Shores, Thomas S. (2007), *Applied linear algebra and matrix analysis*, Springer Science+Business Media, LLC, ISBN 0-387-33194-8.

- Sharipov, Ruslan A. (1996), *Course of Linear Algebra and Multidimensional Geometry: the textbook*, arXiv: math/0405323 (<http://arxiv.org/abs/math/0405323>), ISBN 5-7477-0099-5.
- Gohberg, Israel; Lancaster, Peter; Rodman, Leiba (2005), *Indefinite linear algebra and applications*, Basel-Boston-Berlin: Birkhäuser Verlag, ISBN 3-7643-7349-0.

External links

- What are Eigen Values? (<http://www.physlink.com/education/AskExperts/ae520.cfm>) – non-technical introduction from PhysLink.com's "Ask the Experts"
- Eigen Values and Eigen Vectors Numerical Examples (<http://people.revoledu.com/kardi/tutorial/LinearAlgebra/EigenValueEigenVector.html>) – Tutorial and Interactive Program from Revoledu.
- Introduction to Eigen Vectors and Eigen Values (<http://khanexercises.appspot.com/video?v=PhfbEr2btGQ>) – lecture from Khan Academy
- Hill, Roger (2009). " λ – Eigenvalues" (<http://www.sixtysymbols.com/videos/eigenvalues.htm>). *Sixty Symbols*. Brady Haran for the University of Nottingham.

Theory

- Hazewinkel, Michiel, ed. (2001), "Eigen value" (<http://www.encyclopediaofmath.org/index.php?title=p/e035150>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Hazewinkel, Michiel, ed. (2001), "Eigen vector" (<http://www.encyclopediaofmath.org/index.php?title=p/e035180>), *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Eigenvalue (of a matrix) (<http://planetmath.org/?op=getobj&from=objects&id=4397>), PlanetMath.org.
- Eigenvector (<http://mathworld.wolfram.com/Eigenvector.html>) – Wolfram MathWorld
- Eigen Vector Examination working applet (<http://ocw.mit.edu/ans7870/18/18.06/javademo/Eigen/>)
- Same Eigen Vector Examination as above in a Flash demo with sound (http://web.mit.edu/18.06/www/Demos/eigen-applet-all/eigen_sound_all.html)
- Computation of Eigenvalues (<http://www.sosmath.com/matrix/eigen1/eigen1.html>)
- Numerical solution of eigenvalue problems (<http://www.cs.utk.edu/~dongarra/etemplates/index.html>) Edited by Zhaojun Bai, James Demmel, Jack Dongarra, Axel Ruhe, and Henk van der Vorst
- Eigenvalues and Eigenvectors on the Ask Dr. Math forums: (<http://mathforum.org/library/drmath/view/55483.html>), (<http://mathforum.org/library/drmath/view/51989.html>)

Online calculators

- arndt-bruenner.de (http://www.arndt-bruenner.de/mathe/scripts/engl_eigenwert.htm)
- bluebit.gr (<http://www.bluebit.gr/matrix-calculator/>)
- wims.unice.fr (<http://wims.unice.fr/wims/wims.cgi?session=6S051ABAFA.2&+lang=en&+module=tool/linear/matrix.en>)

Demonstration applets

- Java applet about eigenvectors in the real plane (<http://scienceapplets.blogspot.com/2012/03/eigenvalues-and-eigenvectors.html>)

System of linear equations

In mathematics, a **system of linear equations** (or **linear system**) is a collection of linear equations involving the same set of variables. For example,

$$\begin{aligned} 3x + 2y - z &= 1 \\ 2x - 2y + 4z &= -2 \\ -x + \frac{1}{2}y - z &= 0 \end{aligned}$$

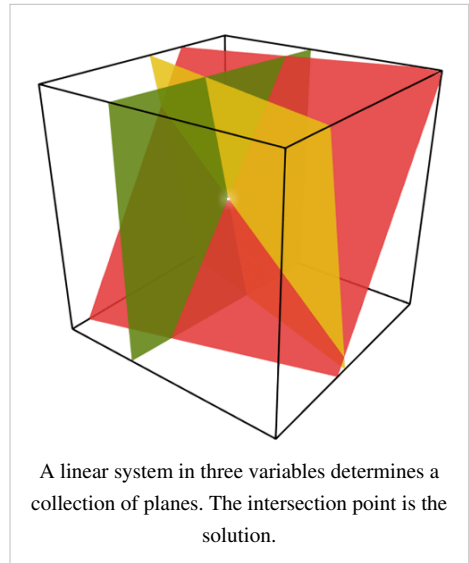
is a system of three equations in the three variables x , y , z . A **solution** to a linear system is an assignment of numbers to the variables such that all the equations are simultaneously satisfied. A solution to the system above is given by

$$\begin{aligned} x &= 1 \\ y &= -2 \\ z &= -2 \end{aligned}$$

since it makes all three equations valid.^[1] The word "*system*" indicates that the equations are to be considered collectively, rather than individually.

In mathematics, the theory of linear systems is the basis and a fundamental part of linear algebra, a subject which is used in most parts of modern mathematics. Computational algorithms for finding the solutions are an important part of numerical linear algebra, and play a prominent role in engineering, physics, chemistry, computer science, and economics. A system of non-linear equations can often be approximated by a linear system (see linearization), a helpful technique when making a mathematical model or computer simulation of a relatively complex system.

Very often, the coefficients of the equations are real or complex numbers and the solutions are searched in the same set of numbers, but the theory and the algorithms apply for coefficients and solutions in any field. For solutions in an integral domain like the ring of the integers, or in other algebraic structures, other theories have been developed, see Linear equation over a ring. Integer linear programming is a collection of method for finding the "best" integer solution (when there are many). Gröbner basis theory provides algorithms when coefficients and unknowns are polynomials. Also tropical geometry is an example of linear algebra in a more exotic structure.



Elementary example

The simplest kind of linear system involves two equations and two variables:

$$2x + 3y = 6$$

$$4x + 9y = 15.$$

One method for solving such a system is as follows. First, solve the top equation for x in terms of y :

$$x = 3 - \frac{3}{2}y.$$

Now substitute this expression for x into the bottom equation:

$$4\left(3 - \frac{3}{2}y\right) + 9y = 15.$$

This results in a single equation involving only the variable y . Solving gives $y = 1$, and substituting this back into the equation for x yields $x = 3/2$. This method generalizes to systems with additional variables (see "elimination of variables" below, or the article on elementary algebra.)

General form

A general system of m linear equations with n unknowns can be written as

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$

$$\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

$$a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m.$$

Here x_1, x_2, \dots, x_n are the unknowns, $a_{11}, a_{12}, \dots, a_{mn}$ are the coefficients of the system, and b_1, b_2, \dots, b_m are the constant terms.

Often the coefficients and unknowns are real or complex numbers, but integers and rational numbers are also seen, as are polynomials and elements of an abstract algebraic structure.

Vector equation

One extremely helpful view is that each unknown is a weight for a column vector in a linear combination.

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

This allows all the language and theory of *vector spaces* (or more generally, *modules*) to be brought to bear. For example, the collection of all possible linear combinations of the vectors on the left-hand side is called their *span*, and the equations have a solution just when the right-hand vector is within that span. If every vector within that span has exactly one expression as a linear combination of the given left-hand vectors, then any solution is unique. In any event, the span has a *basis* of linearly independent vectors that do guarantee exactly one expression; and the number of vectors in that basis (its *dimension*) cannot be larger than m or n , but it can be smaller. This is important because if we have m independent vectors a solution is guaranteed regardless of the right-hand side, and otherwise not guaranteed.

Matrix equation

The vector equation is equivalent to a matrix equation of the form

$$A\mathbf{x} = \mathbf{b}$$

where A is an $m \times n$ matrix, \mathbf{x} is a column vector with n entries, and \mathbf{b} is a column vector with m entries.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

The number of vectors in a basis for the span is now expressed as the *rank* of the matrix.

Solution set

A **solution** of a linear system is an assignment of values to the variables x_1, x_2, \dots, x_n such that each of the equations is satisfied. The set of all possible solutions is called the **solution set**.

A linear system may behave in any one of three possible ways:

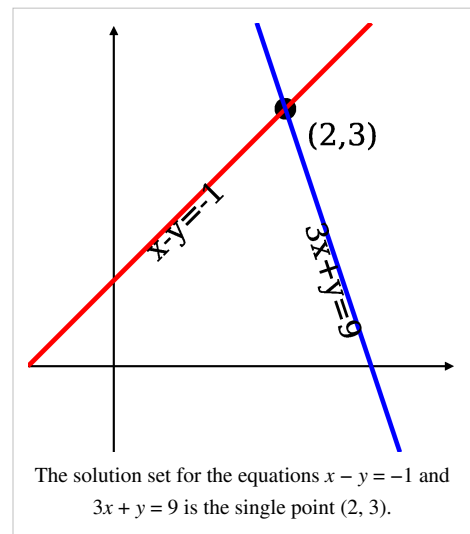
1. The system has *infinitely many solutions*.
2. The system has a single *unique solution*.
3. The system has *no solution*.

Geometric interpretation

For a system involving two variables (x and y), each linear equation determines a line on the xy -plane. Because a solution to a linear system must satisfy all of the equations, the solution set is the intersection of these lines, and is hence either a line, a single point, or the empty set.

For three variables, each linear equation determines a plane in three-dimensional space, and the solution set is the intersection of these planes. Thus the solution set may be a plane, a line, a single point, or the empty set.

For n variables, each linear equation determines a hyperplane in n -dimensional space. The solution set is the intersection of these hyperplanes, which may be a flat of any dimension.

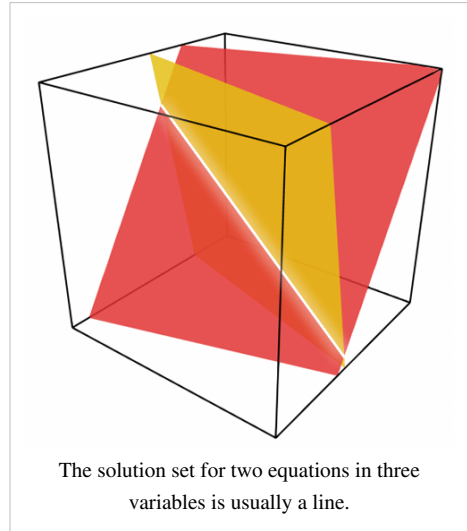


General behavior

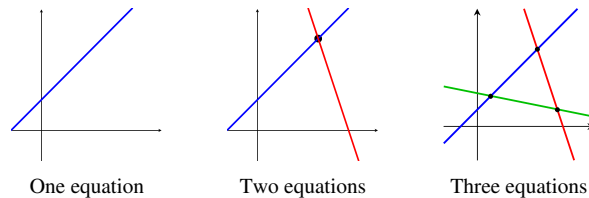
In general, the behavior of a linear system is determined by the relationship between the number of equations and the number of unknowns:

1. Usually, a system with fewer equations than unknowns has infinitely many solutions or sometimes unique sparse solutions (compressed sensing). Such a system is also known as an underdetermined system.
2. Usually, a system with the same number of equations and unknowns has a single unique solution.
3. Usually, a system with more equations than unknowns has no solution. Such a system is also known as an overdetermined system.

In the first case, the dimension of the solution set is usually equal to $n - m$, where n is the number of variables and m is the number of equations.



The following pictures illustrate this trichotomy in the case of two variables:



The first system has infinitely many solutions, namely all of the points on the blue line. The second system has a single unique solution, namely the intersection of the two lines. The third system has no solutions, since the three lines share no common point.

Keep in mind that the pictures above show only the most common case. It is possible for a system of two equations and two unknowns to have no solution (if the two lines are parallel), or for a system of three equations and two unknowns to be solvable (if the three lines intersect at a single point). In general, a system of linear equations may behave differently than expected if the equations are **linearly dependent**, or if two or more of the equations are **inconsistent**.

Properties

Independence

The equations of a linear system are **independent** if none of the equations can be derived algebraically from the others. When the equations are independent, each equation contains new information about the variables, and removing any of the equations increases the size of the solution set. For linear equations, logical independence is the same as linear independence.

For example, the equations

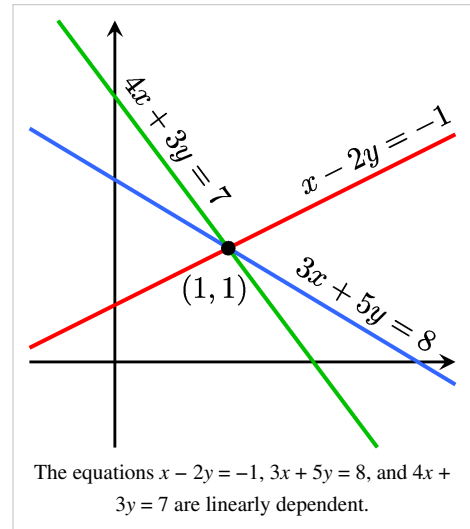
$$3x + 2y = 6 \quad \text{and} \quad 6x + 4y = 12$$

are not independent — they are the same equation when scaled by a factor of two, and they would produce identical graphs. This is an example of equivalence in a system of linear equations.

For a more complicated example, the equations

$$\begin{aligned} x - 2y &= -1 \\ 3x + 5y &= 8 \\ 4x + 3y &= 7 \end{aligned}$$

are not independent, because the third equation is the sum of the other two. Indeed, any one of these equations can be derived from the other two, and any one of the equations can be removed without affecting the solution set. The graphs of these equations are three lines that intersect at a single point.



Consistency

A linear system is **consistent** if it has a solution, and **inconsistent** otherwise. When the system is inconsistent, it is possible to derive a contradiction from the equations, that may always be rewritten such as the statement $0 = 1$.

For example, the equations

$$3x + 2y = 6 \quad \text{and} \quad 3x + 2y = 12$$

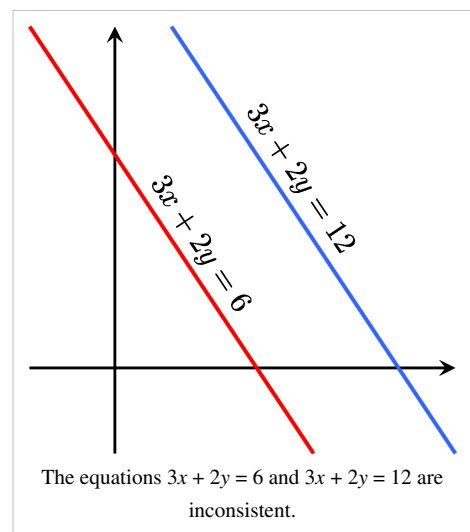
are inconsistent. In fact, by subtracting the first equation from the second one and multiplying both sides of the result by $1/6$, we get $0 = 1$. The graphs of these equations on the xy -plane are a pair of parallel lines.

It is possible for three linear equations to be inconsistent, even though any two of them are consistent together. For example, the equations

$$\begin{aligned} x + y &= 1 \\ 2x + y &= 1 \\ 3x + 2y &= 3 \end{aligned}$$

are inconsistent. Adding the first two equations together gives $3x + 2y = 2$, which can be subtracted from the third equation to yield $0 = 1$. Note that any two of these equations have a common solution. The same phenomenon can occur for any number of equations.

In general, inconsistencies occur if the left-hand sides of the equations in a system are linearly dependent, and the constant terms do not satisfy the dependence relation. A system of equations whose left-hand sides are linearly



independent is always consistent.

Putting it another way, according to the Rouché–Capelli theorem, any system of equations (overdetermined or otherwise) is inconsistent if the rank of the augmented matrix is greater than the rank of the coefficient matrix. If, on the other hand, the ranks of these two matrices are equal, the system must have at least one solution. The solution is unique if and only if the rank equals the number of variables. Otherwise the general solution has k free parameters where k is the difference between the number of variables and the rank; hence in such a case there are an infinitude of solutions. The rank of a system of equations can never be higher than [the number of variables] + 1, which means that a system with any number of equations can always be reduced to a system that has a number of independent equations that is maximum equal to [the number of variables] + 1.

Equivalence

Two linear systems using the same set of variables are **equivalent** if each of the equations in the second system can be derived algebraically from the equations in the first system, and vice-versa. Two systems are equivalent if either both are inconsistent or each equation of any of them is a linear combination of the equations of the other one. It follows that two linear systems are equivalent if and only if they have the same solution set.

Solving a linear system

There are several algorithms for solving a system of linear equations.

Describing the solution

When the solution set is finite, it is reduced to a single element. In this case, the unique solution is described by a sequence of equations whose left hand sides are the names of the unknowns and right hand sides are the corresponding values, for example $(x = 3, y = -2, z = 6)$. When an order on the unknowns has been fixed, for example the alphabetical order the solution may be described as a vector of values, like $(3, -2, 6)$ for the previous example.

It can be difficult to describe a set with infinite solutions. Typically, some of the variables are designated as **free** (or **independent**, or as **parameters**), meaning that they are allowed to take any value, while the remaining variables are **dependent** on the values of the free variables.

For example, consider the following system:

$$\begin{aligned}x + 3y - 2z &= 5 \\ 3x + 5y + 6z &= 7\end{aligned}$$

The solution set to this system can be described by the following equations:

$$x = -7z - 1 \quad \text{and} \quad y = 3z + 2.$$

Here z is the free variable, while x and y are dependent on z . Any point in the solution set can be obtained by first choosing a value for z , and then computing the corresponding values for x and y .

Each free variable gives the solution space one degree of freedom, the number of which is equal to the dimension of the solution set. For example, the solution set for the above equation is a line, since a point in the solution set can be chosen by specifying the value of the parameter z . An infinite solution of higher order may describe a plane, or higher dimensional set.

Different choices for the free variables may lead to different descriptions of the same solution set. For example, the solution to the above equations can alternatively be described as follows:

$$y = -\frac{3}{7}x + \frac{11}{7} \quad \text{and} \quad z = -\frac{1}{7}x - \frac{1}{7}.$$

Here x is the free variable, and y and z are dependent.

Elimination of variables

The simplest method for solving a system of linear equations is to repeatedly eliminate variables. This method can be described as follows:

1. In the first equation, solve for one of the variables in terms of the others.
2. Plug this expression into the remaining equations. This yields a system of equations with one fewer equation and one fewer unknown.
3. Continue until you have reduced the system to a single linear equation.
4. Solve this equation, and then back-substitute until the entire solution is found.

For example, consider the following system:

$$x + 3y - 2z = 5$$

$$3x + 5y + 6z = 7$$

$$2x + 4y + 3z = 8$$

Solving the first equation for x gives $x = 5 + 2z - 3y$, and plugging this into the second and third equation yields

$$-4y + 12z = -8$$

$$-2y + 7z = -2$$

Solving the first of these equations for y yields $y = 2 + 3z$, and plugging this into the second equation yields $z = 2$.

We now have:

$$x = 5 + 2z - 3y$$

$$y = 2 + 3z$$

$$z = 2$$

Substituting $z = 2$ into the second equation gives $y = 8$, and substituting $z = 2$ and $y = 8$ into the first equation yields $x = -15$. Therefore, the solution set is the single point $(x, y, z) = (-15, 8, 2)$.

Row reduction

In **row reduction**, the linear system is represented as an augmented matrix:

$$\left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 3 & 5 & 6 & 7 \\ 2 & 4 & 3 & 8 \end{array} \right].$$

This matrix is then modified using elementary row operations until it reaches reduced row echelon form. There are three types of elementary row operations:

Type 1: Swap the positions of two rows.

Type 2: Multiply a row by a nonzero scalar.

Type 3: Add to one row a scalar multiple of another.

Because these operations are reversible, the augmented matrix produced always represents a linear system that is equivalent to the original.

There are several specific algorithms to row-reduce an augmented matrix, the simplest of which are Gaussian elimination and Gauss-Jordan elimination. The following computation shows Gauss-Jordan elimination applied to the matrix above:

$$\begin{aligned} \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 3 & 5 & 6 & 7 \\ 2 & 4 & 3 & 8 \end{array} \right] &\sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & -4 & 12 & -8 \\ 2 & 4 & 3 & 8 \end{array} \right] \sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & -4 & 12 & -8 \\ 0 & -2 & 7 & -2 \end{array} \right] \sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & 1 & -3 & 2 \\ 0 & -2 & 7 & -2 \end{array} \right] \\ &\sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & 1 & -3 & 2 \\ 0 & 0 & 1 & 2 \end{array} \right] \sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 2 \end{array} \right] \sim \left[\begin{array}{ccc|c} 1 & 3 & 0 & 9 \\ 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 2 \end{array} \right] \sim \left[\begin{array}{ccc|c} 1 & 0 & 0 & -15 \\ 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 2 \end{array} \right]. \end{aligned}$$

The last matrix is in reduced row echelon form, and represents the system $x = -15$, $y = 8$, $z = 2$. A comparison with the example in the previous section on the algebraic elimination of variables shows that these two methods are in fact the same; the difference lies in how the computations are written down.

Cramer's rule

Cramer's rule is an explicit formula for the solution of a system of linear equations, with each variable given by a quotient of two determinants. For example, the solution to the system

$$\begin{aligned}x + 3y - 2z &= 5 \\3x + 5y + 6z &= 7 \\2x + 4y + 3z &= 8\end{aligned}$$

is given by

$$x = \frac{\begin{vmatrix} 5 & 3 & -2 \\ 7 & 5 & 6 \\ 8 & 4 & 3 \end{vmatrix}}{\begin{vmatrix} 1 & 3 & -2 \\ 3 & 5 & 6 \\ 2 & 4 & 3 \end{vmatrix}}, \quad y = \frac{\begin{vmatrix} 1 & 5 & -2 \\ 3 & 7 & 6 \\ 2 & 8 & 3 \end{vmatrix}}{\begin{vmatrix} 1 & 3 & -2 \\ 3 & 5 & 6 \\ 2 & 4 & 3 \end{vmatrix}}, \quad z = \frac{\begin{vmatrix} 1 & 3 & 5 \\ 3 & 5 & 7 \\ 2 & 4 & 8 \end{vmatrix}}{\begin{vmatrix} 1 & 3 & -2 \\ 3 & 5 & 6 \\ 2 & 4 & 3 \end{vmatrix}}.$$

For each variable, the denominator is the determinant of the matrix of coefficients, while the numerator is the determinant of a matrix in which one column has been replaced by the vector of constant terms.

Though Cramer's rule is important theoretically, it has little practical value for large matrices, since the computation of large determinants is somewhat cumbersome. (Indeed, large determinants are most easily computed using row reduction.) Further, Cramer's rule has very poor numerical properties, making it unsuitable for solving even small systems reliably, unless the operations are performed in rational arithmetic with unbounded precision.

Matrix solution

If the equation system is expressed in the matrix form $A\mathbf{x} = \mathbf{b}$, the entire solution set can also be expressed in matrix form. If the matrix A is square (has m rows and $n=m$ columns) and has full rank (all m rows are independent), then the system has a unique solution given by

$$\mathbf{x} = A^{-1}\mathbf{b}$$

where A^{-1} is the inverse of A . More generally, regardless of whether $m=n$ or not and regardless of the rank of A , all solutions (if any exist) are given using the Moore-Penrose pseudoinverse of A , denoted A^g , as follows:

$$\mathbf{x} = A^g\mathbf{b} + (I - A^gA)\mathbf{w}$$

where \mathbf{w} is a vector of free parameters that ranges over all possible $n \times 1$ vectors. A necessary and sufficient condition for any solution(s) to exist is that the potential solution obtained using $\mathbf{w} = \mathbf{0}$ satisfy $A\mathbf{x} = \mathbf{b}$ — that is, that $AA^g\mathbf{b} = \mathbf{b}$. If this condition does not hold, the equation system is inconsistent and has no solution. If the condition holds, the system is consistent and at least one solution exists. For example, in the above-mentioned case in which A is square and of full rank, A^g simply equals A^{-1} and the general solution equation simplifies to $\mathbf{x} = A^{-1}\mathbf{b} + (I - A^{-1}A)\mathbf{w} = A^{-1}\mathbf{b} + (I - I)\mathbf{w} = A^{-1}\mathbf{b}$ as previously stated, where \mathbf{w} has completely dropped out of the solution, leaving only a single solution. In other cases, though, \mathbf{w} remains and hence an infinitude of potential values of the free parameter vector \mathbf{w} give an infinitude of solutions of the equation.

Other methods

While systems of three or four equations can be readily solved by hand, computers are often used for larger systems. The standard algorithm for solving a system of linear equations is based on Gaussian elimination with some modifications. Firstly, it is essential to avoid division by small numbers, which may lead to inaccurate results. This can be done by reordering the equations if necessary, a process known as *pivoting*. Secondly, the algorithm does not exactly do Gaussian elimination, but it computes the LU decomposition of the matrix A . This is mostly an organizational tool, but it is much quicker if one has to solve several systems with the same matrix A but different vectors \mathbf{b} .

If the matrix A has some special structure, this can be exploited to obtain faster or more accurate algorithms. For instance, systems with a symmetric positive definite matrix can be solved twice as fast with the Cholesky decomposition. Levinson recursion is a fast method for Toeplitz matrices. Special methods exist also for matrices with many zero elements (so-called sparse matrices), which appear often in applications.

A completely different approach is often taken for very large systems, which would otherwise take too much time or memory. The idea is to start with an initial approximation to the solution (which does not have to be accurate at all), and to change this approximation in several steps to bring it closer to the true solution. Once the approximation is sufficiently accurate, this is taken to be the solution to the system. This leads to the class of iterative methods.

Homogeneous systems

A system of linear equations is **homogeneous** if all of the constant terms are zero:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= 0 \\ \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= 0. \end{aligned}$$

A homogeneous system is equivalent to a matrix equation of the form

$$A\mathbf{x} = \mathbf{0}$$

where A is an $m \times n$ matrix, \mathbf{x} is a column vector with n entries, and $\mathbf{0}$ is the zero vector with m entries.

Solution set

Every homogeneous system has at least one solution, known as the **zero solution** (or **trivial solution**), which is obtained by assigning the value of zero to each of the variables. If the system has a non-singular matrix ($\det(A) \neq 0$) then it is also the only solution. If the system has a singular matrix then there is a solution set with an infinite number of solutions. This solution set has the following additional properties:

1. If \mathbf{u} and \mathbf{v} are two vectors representing solutions to a homogeneous system, then the vector sum $\mathbf{u} + \mathbf{v}$ is also a solution to the system.
2. If \mathbf{u} is a vector representing a solution to a homogeneous system, and r is any scalar, then $r\mathbf{u}$ is also a solution to the system.

These are exactly the properties required for the solution set to be a linear subspace of \mathbf{R}^n . In particular, the solution set to a homogeneous system is the same as the null space of the corresponding matrix A .

Relation to nonhomogeneous systems

There is a close relationship between the solutions to a linear system and the solutions to the corresponding homogeneous system:

$$A\mathbf{x} = \mathbf{b} \quad \text{and} \quad A\mathbf{x} = \mathbf{0}.$$

Specifically, if \mathbf{p} is any specific solution to the linear system $A\mathbf{x} = \mathbf{b}$, then the entire solution set can be described as

$$\{\mathbf{p} + \mathbf{v} : \mathbf{v} \text{ is any solution to } A\mathbf{x} = \mathbf{0}\}.$$

Geometrically, this says that the solution set for $A\mathbf{x} = \mathbf{b}$ is a translation of the solution set for $A\mathbf{x} = \mathbf{0}$. Specifically, the flat for the first system can be obtained by translating the linear subspace for the homogeneous system by the vector \mathbf{p} .

This reasoning only applies if the system $A\mathbf{x} = \mathbf{b}$ has at least one solution. This occurs if and only if the vector \mathbf{b} lies in the image of the linear transformation A .

Notes

[1] Linear algebra, as discussed in this article, is a very well established mathematical discipline for which there are many sources. Almost all of the material in this article can be found in Lay 2005, Meyer 2001, and Strang 2005.

References

Textbooks

- Axler, Sheldon Jay (1997), *Linear Algebra Done Right* (2nd ed.), Springer-Verlag, ISBN 0-387-98259-0
- Lay, David C. (August 22, 2005), *Linear Algebra and Its Applications* (3rd ed.), Addison Wesley, ISBN 978-0-321-28713-7
- Meyer, Carl D. (February 15, 2001), *Matrix Analysis and Applied Linear Algebra* (<http://www.matrixanalysis.com/DownloadChapters.html>), Society for Industrial and Applied Mathematics (SIAM), ISBN 978-0-89871-454-8
- Poole, David (2006), *Linear Algebra: A Modern Introduction* (2nd ed.), Brooks/Cole, ISBN 0-534-99845-3
- Anton, Howard (2005), *Elementary Linear Algebra (Applications Version)* (9th ed.), Wiley International
- Leon, Steven J. (2006), *Linear Algebra With Applications* (7th ed.), Pearson Prentice Hall
- Strang, Gilbert (2005), *Linear Algebra and Its Applications*

Article Sources and Contributors

Mean Source: <https://en.wikipedia.org/w/index.php?oldid=591551746> *Contributors:* 02barryc, 123candy, 128.195.169.xxx, 165.123.179.xxx, 16@r, 212.153.190.xxx, 213.253.39.xxx, Shin3, 7121989, 78.26, A314268, A930913, APT, Adambro, Adamjlsund, Addihockey10, Ahruman, Aitias, Andy Liefthing, Alansohn, Ale jrb, Alexwuv, Alksentrs, Allen3, Allens, Alperen, Altaïr, Altenmann, Amahoney, Amirab, Ammiragiodor, Ann Stouter, Anna Lincoln, Antandros, Arakunem, Arcadian, Arithmetic, Army1987, Armaugir, Arthena, Aruton, Ategeler, AubjJR., Avenue, Avoided, Awaterl, AxelBoldt, B4hand, Babycruzrocks, Bagatelle, Bakirfarhi, Banes, BarretB, Bart133, BenFrantzDale, Bequal, Berland, Bfigura, Bjcairs, Blehfu, BlueEditor, Bobble2, Boblet53, Bost192, Bobrayner, Bongwarrior, Brossow, C'est moi, CMW275, COMPFUNK2, CRGreathouse, CWii, CanadianPenguin, CardinalDan, Carmen56, Casper2k3, Catch2011, Catgut, Chickchick19, Cmacpher, Cmglee, Connelly, Conversion script, Cooolaaron88, Corecirculator, Cronholm144, CroydThoth, Czrussian, Cubs Fan, Cybercobra, DVD R W, DaC04, Dandy, DarkFalls, Davidruben, Dbfirs, DearPrudence, Dellidot, Den fjättrade ankan, DerBorg, DesolateReality, Dethme0w, Diegotorquemada, Dietcokelime2006, Dingo1729, Discodotron, Discospinster, Dktiwari99, Dmcq, Donarreiskoffer, Donner60, DrMicro, Drmies, Duoduoduo, Dysepsion, ERcheck, ESKog, Ear121, EddtheGr8one, Ejj2007, El C, Epr123, Eubulides, EuroCarGT, Falcon8765, Faradayplank, FastLizard4, Fgnievinski, Fieldday-sunday, Fireice, Flyer22, Footwarrior, Fortdj33, Freedomlinux, Freshneesz, Funkystuff267, Funnybunny123, G716, GTBacchus, Gary King, Gene Nygaard, GeoWriter, Ghazer, Giftlite, Ginsuloft, Gogo Dodo, Graham87, Grant meads, Grim23, Groengras, Gul e, Gurch, Gwernol, Gyrofrog, Götz, Hadleywickham, Hamoudafag, HenningThielemann, Henrygb, Hithishal, Hobartimus, Hongoui, Honker04, Hossein-amidi, Hüm, Hugozam, HumbleGod, Hyposifif, I am One of Many, IRP, Icairns, Impdog, Infovarius, Into The Fray, Iridescent, Ivokabel, J roc69, J.delanoy, Jackfork, JaeDyWolf, Jdm, Jeffrey Mall, Jem1299, Jitse Niesen, Jock, John Cline, Johmeneghini, Jok2000, Jonathan de Boyne Pollard, Jschnur, Kaimbridge, Katalaveno, Kazvorpap, Kbdank71, Kevin Lamoreau, KinaseD, Kjetil1001, Kkbairi, Kumioko (renamed), L Kensington, Lapqzmlapq, Learner505, Leondz, LibLord, Linkracer, Loonymonkey, Lotje, Lugia2453, Luna Santin, MacMed, Maksim-e, MarkSweep, Marshall Williams2, Martarius, Marymoo171717, MaterialsScientist, Matty.007, Maxem, McSly, Mebiancame, Mediran, Megaman e m, Melcombe, Mets501, Michael Hardy, Mike Segal, Mike20579, Mikeblas, Mild Bill Hiccup, Mmortal03, Mpj17, MrOllie, Mschindwein, Mwtoews, NawlinWiki, Nburden, Neko-chan, Neonxing23, Neptune5000, Norazoye, NotAnonymous0, Nwerneck, Octahedron80, Oleg Alexandrov, Omnipaedista, Oore, Op47, Orphan Wiki, Oxymeron83, Ozob, ParticleMan, Partycos, Patrick, Paul August, Pepper, Persian Poet Gal, PeterSymonds, Pharaoh of the Wizards, Philip Trueman, Philomatholitic, Piano non troppo, Pinethicket, Polar, Porejide, Pouply, Pseudomonas, Pt, Qwfp, R. S. Shaw, RJaguar3, RL0919, Rachel jean, Redvers, Reza1615, Riahmare815, Rich Farmbrough, Riri 16, Rjohson92, Ronhones, Ronz, Ryan Vesey, Salix alba, Sam Derbyshire, Samwb123, Sanford, Sanjiv swarup, Sbfw, Schmo, Scohous, Sct72, Seba5618, Secret, SecretLondon, Sempert, Shirulashem, Shorelander, Silly rabbit, Silvrous, Sixro, Skizzik, SkyWalker, Slamb, Skloding Man, Smithpfit, Snowfoll, Some jerk on the Internet, Some poser, Sp3000, Spazure, Spert0469, Sriharsh1234, Srleffler, Stephenb, Suffusion of Yellow, Sunderland06, Sven Manguard, Synchronism, Syrthiss, Slawomir Biały, TPK, TacomaZach, Tanaats, Tangerines, Taransingh63, Teklund, Temerster, Tentinator, Tero, Tex, Thatotherperson, The Rumbling Man, The Red, The Thing That Should Not Be, ThePolisime, Thomas.Hedden, Tide rolls, TimothyBrilliantEE, Tom-, Tow, Trevor MacInnis, Vadith, Vanished user uih38riiuv4hJsd, Vcpandya, Vegaswikian, Victoriaplummer, Vrenator, Verkoopel, Vsb, Waldir, Wbm1058, Weblent101, Whatfjg, Whywhenwhohow, Widr, WikiPuppies, Willking1979, With goodness in mind, Wjejskenewr, Woogee, Wyatt915, Xyb, Yono, Youtookitback9, Zacharrington, Zoolium, Zvika, 712 anonymous edits

Median Source: <https://en.wikipedia.org/w/index.php?oldid=593249980> *Contributors:* 16@r, 4johnny, 777sms, AGToth, Acebulf, Adam78, Adamjlsund, Afa86, Airplaneman, Alejo2083, Almuayyad, AlphaEta, Ancheta Wis, Andrew c, Andrewman327, Andrewpmk, Anrnsusa, Antandros, Apanag, Arcadian, Art LaPella, Arthena, Atehetkos, Atlant, Avenged Eightfold, Avjoska, Avowedun, AxelBoldt, B4hand, Baccyak4H, Bender235, Benwing, Bfinn, Bhudson, Billinghurst, Bjcairs, Blanchardb, Blindman shady, Bluap, Brain40, Brianjd, BrotherE, Bshort, Bth, Capricorn42, CharlotteWebb, Chire, ChrisGualtieri, Clarkj12, Cmglee, ComiC1, Connelly, Conversion script, Cryptor3, Cvaneg, Cybercobra, Cybersavior, DL5MDA, DVdm, Daniel.Cardenas, Dannyyoung97, DarkThunder2, David Epstein, Deljr, Dcoetzee, Den fjättrade ankan, Derek farn, Dick Beldin, Dima373, Diomidis Spinellis, Dirkb, Discospinster, Donarreiskoffer, Doublebellephonium, DrMicro, Dreadstar, Dricherby, Drphilharmonic, Dswallen, Dtuinhof, Duoduoduo, Epr123, Eric Kvaalen, Eseiio, Evil saltine, Explicit, Eyliu, Fangz, Fgnievinski, Foxtrot620, FreplySpang, Fruggo, G716, GIScope, Gatewaycat, Georg-Johann, Giftlite, Gilly 54, Glane23, Gomm, Graham5571, GreenGourd, Grossu, Gwernol, Hadleywickham, Haham hanuka, HamburgerRadio, Hammersoft, Hamtechperson, Heimstern, Henrygb, Herbee, Hiram 99, Hottentot, Hut 8.5, IMpbt, Impdog, Incnis Mersi, Insanity Incarnate, J Hill, J.delanoy, JForget, JackOL31, Janneok, Jason Quinn, Jawavizard, Jesdisciple, Jfpierce, Jitse Niesen, Jkln, Jobin RV, JodyB, Jose Ramos, JoshuaSchaeffer, Katalaveno, Kateshortforbob, Kiefer.Wolfowitz, Kpjias, Kuru, Kurzon, Kwamikagami, LGF1992UK, Lambiam, LeaveSleaves, Lirion, LivingFont, Luigifan, Lythanphu, MER-C, MacMed, Maddiel, Male1979, Manop, Maple, Marco Pellegrino, MarkSweep, Martarius, MaterialsScientist, Mathstat, MattGiuca, Mclid, Melcombe, Mentifisto, Michael Hardy, Michaelas10, Mikhail Ryazanov, Minesweeper, Mishraskneuh, Mormegil, MrOllie, MusikAnimal, Mysterious Whisper, Nakon, Nascar1996, Navit.mahajan, Nbarth, Ncmvocalist, Ndenison, NerdyPunk2ML, Nishkid64, Nixdorf, Noctibus, Norazoye, Obradovic Goran, Octahedron80, Op47, Oskar Sigvardsson, Pablo Alcayaga, Palnatoke, Pamri, Patrick, Paul August, Penguinerr121, Perubaby44, Peter.vanroose, Pinethicket, Policron, Poulntey02, Quantling, Qwertusy, Qwfp, RL0919, Radiojon, Ratonhaketon, RexNL, Rich Farmbrough, Robma, Roozbeh, RoseParks, Rumping, RyanCross, SURIV, Salix alba, Sameersingh, SanjivBhatia, Schmidtm, Seresin, Sexysexy555, Sfdan, Shannon1 (usurped3), Shmget, Signalhead, SlamDiego, Sligocki, SoSaysChappy, Sobreira, Squids and Chips, Steinsky, Stemonitis, Spasha, Struway, Surfgr8r13, Swaroopch, Slawomir Biały, TPK, Tentaclestraw3, Thatguyflint, The Thing That Should Not Be, Thnidu, Thomas Tvleren, Thumani Mabwe, Tobias Bergegam, Tomi, Triskell, Trusilver, Tsirel, U3002, Urhixidur, Vishnava, Wavelength, Wbm1058, WejjiBaikeBianji, Well, girl, look at you!, Wetman, Whywhenwhohow, WikHead, WikiDao, WikipedianMarlith, Wildscop, Woleball, XXJASHANXX, Zundark, Zvika, Zyxoas, ^demon, 512 anonymous edits

Mode (statistics) Source: <https://en.wikipedia.org/w/index.php?oldid=589847646> *Contributors:* ABF, AGToth, Adam78, Adamjlsund, Aitias, Alansohn, Anonymous Dissident, Arcadian, Arcandam, Arda Xi, Arowana1997, Arthena, Arthur Rubin, Ash CUFC, Avenue, AzureCitizen, Belg4mit, BenFrantzDale, Bender235, Blanchardb, CMBJ, Ched, Cmglee, Cornetysheuse, Cretog8, Cruccione, Cybercobra, Cyrillic, DBigXray, DEMcAdams, DTPolet, Danski14, DavidCBryant, DeadEyeArrow, Den fjättrade ankan, Diegotorquemada, Diomidis Spinellis, Donner60, DrMicro, Eclectics, EdColbert, Ehrenkater, El C, Epr123, Eric-Wester, Foxtrot620, Freshneesz, Fusionmix, G716, George The Dragon, Giftlite, Giro720, Gombo, Hajatvur, Henrygb, Hisabness, Hogwarts boy, Horkanomonno, Horpana, Hugesim, I dream of horses, IRP, IloveResearching234, Inferno, Lord of Penguins, Isheden, J.delanoy, JForget, Jacob Lundberg, Jamelan, James Cantor, Jitse Niesen, John Cline, Jxramos, Jóna Þórunn, Kartikmohta, Katoa, Kinema, Lambiam, Lmlms44, Lonestarnot, MacMed, Macterra, Maggyero, Manop, Marcusmax, MarkSweep, Martarius, Matthew Yeager, MattieTK, Mclid, Melcombe, Memming, Mhaitiam.shammaa, Mormegil, MrOllie, Musiphil, Mythmon, Nimbusania, Oily150, Overmind 900, Paintman, Paul August, Pdcurry, PhantomTech, Philip Trueman, Phynicen, PinkSbat, Pmanderson, Policron, Qwfp, R'n'B, R000t, RainbOWlight, Rich Farmbrough, Richard001, Rrurke, Rumping, Sameer0s, Seaphoto, Simeon.mattes, Slakr, Starlady14, Super48paul, Talgalili, Tgeairn, That Guy, From That Show!, The High Fin Sperm Whale, The Thing That Should Not Be, Thingg, Thumani Mabwe, Tide rolls, Ulmanor, Versus22, Victoriaplummer, Wavelength, Whywhenwhohow, Widr, Wikieditor06, Zheric, 292 anonymous edits

Variance Source: <https://en.wikipedia.org/w/index.php?oldid=593274094> *Contributors:* 16@r, 212.153.190.xxx, 28bytes, ABCD, Aastrup, Abramjackson, AbsolutDan, Accretivehealth, Adamjlsund, Adonijahowns, Adpete, Afa86, Ahoersteimer, Alai, Albmont, Alex756, AmiDaniel, Amir Alev, Amircrypto, Anake.xii, Anameofmyveryown, Andre.holzner, AndzejGlo, Angela, Animun, AntiVMan, Anuphysicguy, As530, Auntof6, Awickert, Baccyak4H, Baptiste R, BarroColorado, Bart133, Bcc32, BenFrantzDale, Bender235, Blaisorblade, Blotwell, Bnjj, Bobo The Ninja, Borgx, Brandon Moore, Brian Sayrs, Bryan Derksen, Brzak, Btyner, Callanec, CanDo, Casey Abell, Cazort, Centrx, Cfp, Cgsguy2, Cherkash, Chire, ChrisGualtieri, Coffee2theorems, Compassghost, Conversion script, Coppertwig, Cremepuff222, Cruise, Cryptomatt, Cumulonix, Cybercobra, DRE, Darko.veberic, Davi393, DavidCBryant, Davwillev, Deljr, Dearlighton, Den fjättrade ankan, Diophantus, Disavian, Discospinster, DoctorW, Docu, Double Blind, DrMicro, Duncan, Duncharris, Duoduoduo, Dylan Lake, Efe485, Ehrenkater, Elgreengeeto, Emraherr, EnJx, Eric-Wester, Eric.nickel, Erxnmedia, EtudiantEco, Eykanal, Fangorn-Y, Fgnievinski, Fibonacci, Foam bubble, François Robere, Freddie, G716, Gap, Garamatt, George Ponderevo, Gh02t, Giftlite, Gjshisha, Glimz, Graft, Guanaco, Gurch, Gzkn, Hao2lian, Happy-melon, Hede2000, Herath.sanjeewa, Het, Hgberman, Ht686rg90, Hu12, Hulk1986, I am not a dog, Inezz40, Inter16, Isaac Dupree, Isj, J.delanoy, JackSchmidt, Jackzhp, Jasondet, Jesse V., Jfessler, Jheald, Jheiv, Jmath666, Joeejc, Joepwijers, JohnBlackburne, Johnny Au, Josh Cherry, Jsharppminor, J688, Juha, JulesEllis, Julia Abril, Junkinbomb, Justin W Smith, Jutta, Kastchei, Katzmik, Keenan Pepper, Keilana, Kevmitch, Kiatdd, Kiefer.Wolfowitz, Kiril Simeonovski, Kik206, Kstarsinic, Kurykh, Kymacpherson, LOL, Lambiam, Larry_Sanger, LeaW, Lilac Soul, Linas, Liyuxuan1412, Madprog, Mandarax, Marek69, MarkSweep, MaterialsScientist, Mathstat, Matt Cook, Matthew.daniels, Maxi, Mbloore, McKay, Mebden, Philip Los Indios, Melcombe, Mgreembe, Michael Hardy, Michel M Verstraete, Mike Rosoft, Mikhail Ryazanov, Mjg3456789, MrOllie, Msanford, MsciffyThree, Mwtoews, Natalie Erin, Nbarth, Nevillerichards, Nicogla, Nijdam, Nixphoeni, Nojhan, Notedgrant, O18, Oleg Alexandrov, Orphan Wiki, Ottawa4ever, Paolo.dL, Paresnah, Patrick, Paul August, Paul Pogonyashev, Paul2520, Paulust2002, PerfectStorm, Pgan002, Phantomsteve, Philg88, Phoenix00017, Physicsdz, Pichote, PimRijkee, Pinethicket, Piotrus, Pmanderson, Pokipsy76, Psychlohexane, Qwfp, Raffamaden, Ranger2006, Rbj, Recognizance, Rich Farmbrough, RobertCup, RobinK, Robinh, Roelloonen, Romanski, Rtc, SD5, Salix alba, Sanchem, SchfiftyThree, SereneStorm, Shoofdeath, Shoyer, Shreevatsa, SimonP, Sinverso, Sirmumberguy, Skbkekas, Sligocki, Slon02, Smartcat, Snafflekid, Spinality, Spoon!, Spasha, Syz2, Talldave, TedPavlic, That Don Guy, The Thing That Should Not Be, Thecheyskid, Thermochap, Thesilverball, Thomag, Tide rolls, Tilo1111, Tim Starling, Timlutre, TomYHChan, Tomi, TomyDuby, Tpsreynolds, Unamofa, Unyoyega, Vaughan Pratt, Voidxor, Waldir, Wavelength, Wikidrone, Wikiqiu11, Wikomidia, William Graham, Wmahan, WordsOnLitmusPaper, Wykpydyda, Yamamoto Ichiro, Yeulay, Zhouji2010, Zippanova, Zirconsot, Zundark, Zven, Bopic Πραξα, 603 anonymous edits

Standard deviation Source: <https://en.wikipedia.org/w/index.php?oldid=593123295> *Contributors:* 1exec1, 83440m, A.amitkumar, AJR, AManWithNoPlan, Abalter, Aberglaube, Abscissa, AbsolutDan, Abtin, Ace of Spades, Adamjlsund, Addshore, Adi4094, Admissions, Aednichols, Aeriform, Afa86, Alansohn, Ale jrb, Alex.g, Alexandrov, AllCluesKey, Allesia67, Alvinvc, Amahoney, Amatulich, Ameid, Amitch, Amithell125, Amorin Parga, Anameofmyveryown, Andraide, Andre Engels, Andres, AndrewWTaylor, Andy Marchbanks, Andycjp, Andysor, AngelOFSadness, Anomie, Anon lynx, Anonymous Dissident, Anonymous editor, Anonymousuk2015, Anrnsusa, Ansa211, Anwar saadat, Arbitrarily0, Arnaugir, Aroundthewayboy, Arthur Rubin, Artichoker, Artorius, Asaba, Ashawley, Ashiabor, Astatine211, Atehetkos, AugPi, AxelBoldt, Bart133, Bdesham, Beefyt, Beetstra, Behco, Beland, BenFrantzDale, Bgwhite, Bha100710, BiT, Billgordon1099, Blah314, Blehfu, Bo Jacoby, Bobo192, Bodnotbud, Brianga, Brutha, BryanG, Bsdmide, Btyner, Buchanan-Hermit, Bulgarotanon, Butcheries, CJLL, Wright, CRGreathouse, CSWarren, CWii, CYD, Calabel1992, Calculator1000, Calvin 1998, CambridgeBayWeather, Captain-n00dle, Carmichael, Cathardic, CcDbasmile735593, Cenkuyan, Ceyockey, Coffee Matthews, Chafacter, Chillwithabong, Chris the speller, ChrisFoneten13, ChrisGualtieri, Chrism, Christopher Parham, Chrysoybyn, Cjarrodsnow, Ck lostword, Clemwang, Cmicahel, Cuffee, Coffee2theorems, Conversion script, Coppertwig, Corrigann, Crazy Boris with a red beard, Crisófilax, Cryptic C62, Cutler, DARTH SIDIOUS 2, DRHagen, DRTLtbrg, DVD R W, DVdm, Daev, Danger, DanielCD, Danielb613, Danski14, Darko.veberic, Dave6, DavidLeighEllis, DavidMcKenzie, DavidSJ, DavidWBrooks, Davidkazuhiro, Davidwbulger, Dcoetzee, Ddiashn, Ddofborg, Ddr, Decayintodust, DeeDeeKerby, Dekuntz, Dellidot, Den fjättrade ankan, Denisarona, Dennyhelper, DerHexer, Derekleungtszhei, Dgw, Dhanya139, Dick Beldin,

DieterVanUytvanck, Diomidis Spinellis, Dirkbb, Discospinster, Dmcq, Dmr2, Doctorambient, DomClea, Dominus, Donner60, Dr.alaaagad, DrMicro, Drappel, Dreslough, Duoduoduo, Dycedarg, Dylan Lake, Dymitr, EJM86, EPublicRelationsMT, Earth, Eb Oesch, Economist2007, Eggriffin, Egan, Elaragil, Ellisbe, Emerah, Endobrendo, Enigmaman, Epr123, Eric Olson, Erutuon, Esrever, Eve Teschelmacher, Everyking, Excirial, Falcon8765, Falcon9x5, Falta00, Felixriving, Felnievinski, FilipeS, Flamurai, Forlornturtle, Forty two, Frehley, Frencheigh, Fslser, Furrykef, Fæ, G.engelstein, G716, Gabbe, Gail, Gary King, Gatoclass, Gauge, Gauravm1312, Gemini1980, Geneffects, George Drummond, George Ponderevo, Gerriet42, Gherghel, Giftlite, Gilliam, Gingemonkey, Giraffedata, Gjshisha, GlassCoBra, Glen, Gogowitsch, Gouveia2, Graham87, Greenstray, Greg L, GregWooleedge, Gurch, Gvanrossum, Gyro Copter, Gzkn, Gökhan, HaakonHjortland, Hadleywickham, Haham hanuka, Haizum, HalfShadow, Harmil, Hatashi, Hawaiian1717, Hede2000, Heezy, Helix84, Hellknorz, HesterSheng, Hgberman, Hgrenbor, Hilgerdenaar, Hu12, Hut 8.5, Iccaldwell, Imagine Reason, Imeriki al-Shimoni, Imomyabcs, Intangir, Iridescent, Isaac Dupree, Isis, Isomorphic, JustinPop, Izno, JA(000)Davidson, JForget, JNW, JRBrown, Jacob grace, JadelnOz, Jake04961, Jmamala, James12345, Jamned, Janderk, JanetteDoe, Jcw69, Jean15paul, Jeremy68, Jeremykemp, Jfjtz, Jim.belk, Jim.henderson, JjinJian, Jmoorhouse, Jni, Joerite, John, John Newbury, John11235813, JohnBlackburne, JohnCD, JorivS, Jprg1966, Jratt, Jennie, Jts10101, JulioSergio, Justanyone, Justinep, KJS77, KKOolstra, Kainaw, Kastchei, Kbolino, Kelvie, Khalanil, Khazar2, Khunglongcon, Kidakaka, Kingpin13, Kingturtle, Kirachinmoku, Kiril Simeonovski, Kjtbo, Knkv, Knutux, Krinkle, Krishna91, Kungfuadam, Kuratowski's Ghost, Kuru, Kvg, Kwamikagami, Kyle824, LGW3, LWG, Lambiam, Larry_Sanger, Ldm, Learning4ever, LeaveLeaves, Legare, Legoman 86, Lethe, Lgauthie, Lilac Soul, LizardJr8, Loadmaster, Loodog, Lord Jubjub, Lucasgw8, Lugia2453, Luna Santin, LysolPionex, M1arvin, M2Ys4U, M360 Real, MBisanz, MCEpek, MER-C, MONGO, MZMcBride, Macronencer, Madir, Madoka, MagneticFlux, Magnus Bakken, Malo, Mapley, Marcos, Marek69, MarkSweep, Markhebnr, Markkawika, Markpravda, Marokwitz, MaterialsScientist, Matthew Yeager, Maurice Carbonaro, Mavemp, Mbloore, Mbroshi, Mbweissman, McKay, Mconson, Mdann52, Mehtaba, Melchoir, Melcombe, Mercuryeagle, Mesoderm, Methecoolude, Mets501, Mmgiganteus 1, Mhinckley, Miaow Miaow, Michael Hardy, Miguel.mateo, Mike Rosoft, Mikhail Dvorkin, Richard001, Rickyfrizzlebum, Robo Cop, Rocketrod1960, Rompe, Ronz, Rose Garden, Rsey267, Mjroyster, Moeron, Mollerup, Monteboi1, Mooli, MrOllie, Ms2ger, Msm, Mud4t, Munckin, Murb-, Mwtoews, NHRHS2010, Nakon, Nathandean, NawlinWiki, Nbarth, Nectarflowed, NeilRickards, Neo Poz, Neudachnik, Neuralwarp, Neutrality, Ngoddard, Niallharkin, Nigholth, Noe, Nonagonal Spider, Norazoey, Normy rox, NorwegianBlue, Not-just-yeti, NotAnonymous0, Novalis, Nwbeeson, O.Kosulowski, O18, Octahedron80, Ocvaeils, Oddbdox, Ohl4xoaS, Oleg Alexandrov, Oliphant, Spiffly, Ststone, Statistafactions, Stemonitis, Stepheng3, Stevvers, StewartMH, Storkk, Stpasha, Sufusion of Yellow, Surachit, Susurus, Svick, Swat671, Swcurran, SwisterTwister, THEN WHO WAS PHONE?, Takanoha, Taxman, Tayste, TedE, Tempodivalse, Thadius856, The Thing That Should Not Be, The sock that should not be, Thingy, Thingstofollow, Thomag, ThomasNichols, ThomasStrohman, Thr4wn, Tide rolls, Time9, Titoxd, Tiroche, Tom harrison, Tomi, Tompa, Torsionalmetric, Tosayit, Tpradbury, TradingBands, Triwbe, Urdutexit, Useight, Ute in DC, V DESAI UK, Vaughan Pratt, Verbum Veritas, Versus22, Vice regent, VictorAnyakin, Vinodkumar2, VladimirReshetnikov, W4chris, Waggars, Warniats, Wavelength, Wayne Slam, Wbm1058, Widr, Wikipelli, Wildingd, William Avery, Winchelsea, Winsteps, Wmaham, Wolfkeeper, Wongkld, Woodd, Wykpydyda, X-Fi6, Yaara dildaara, Yachtsman1, Yamaguchi先生, Ylloh, Yochai Twitto, Zafiroblue05, Zenkat, Zenomax, Zhieanag, Zigger, Zvika, زرشك, 1714 anonymous edits

Coefficient of variation *Source:* <https://en.wikipedia.org/w/index.php?oldid=590517123> *Contributors:* AndreasWittenstein, Arunsingh16, Avoided, Berland, BobKawanaka, Btyner, Charles Matthews, Chenopodiaceous, Chrike, Chris Capoccia, Clhtnk, Cyocum, Den fjättrade ankan, Denisarona, Duoduoduo, Fgnievinski, Fram, GL, Gareth Jones, Giftlite, Graham87, Guo, Jack Greenmaven, Jim1138, Jmkin dot com, Johannes Forkman (SLU), Jowa fan, Kinaro, Knuckles, Lamro, Loom91, MartinSpacek, Melcombe, Memeri, Michael Hardy, Nabla, Nbarth, Norazoey, O18, Omegium, Pengortm, PeteLaud, Pinethicket, Qwfp, R'n'B, Sdurf, Shoefly, Strafpeloton2, Tarbo, Tentinator, The Last V8, Tkreuz, Vohuman, Wurzel33, Zenkat, 91 anonymous edits

Skewness *Source:* <https://en.wikipedia.org/w/index.php?oldid=590090038> *Contributors:* 1exec1, 2D, 777sms, ANONYMOUS COWARD0xCODE, Accretivehealth, Alaeix, Amarsesh, Audriusa, Awickert, AxelBoldt, BabbaQ, BenFrantzDale, Bjoram11@yahoo.co.in, BrotherE, Calbaer, Cherkash, Chris the speller, Christopher Mahan, Christopherlin, Cmglee, CommonsDelinker, Conversion script, Courcelles, Cruise, DGX, Dale Arnett, David Haslam, Den fjättrade ankan, Donarreiskoffer, DrMicro, Duoduoduo, Dzertyx, Econotechie, Efender, Eliezg, Empty Buffer, Eric Kvaalen, Fadiwiki, FrankDev, Freelance Intellectual, G716, Giftlite, Ginsuloft, Gjshisha, Graham87, Grrrats, Hankwang, Hargrw, Henrygb, Hve, Illia Connell, Itamblyn, Ivan rove, Ixf64, Jacob Lundberg, Jauerbaek, JavOs, Jdm64, Jim.henderson, Jitse Niesen, Jksad, Jmarjch, Joxemai, Jneill, Jusses2, Justcop4, Jérôme, Kappa, Kareekacha, Kodiologist, Lamro, Lebha, Lingwit, Lynxoid84, MMKO, Madheros88, MarkSweep, Martial75, MartinPoulter, Mathstat, Mcorazao, Melcombe, Memming, Michael Hardy, MrOllie, Muhandes, Myhvac, Nbarth, PBH, Peter Hiemeyer, Petitjeanmichel, Petr C., Pgdfor, Piotrus, PipingHotSoup, Policron, Porejide, Pps, Qwfp, Ramin Nakisa, Rjwilmsi, Rodolfo Hermans, Rumping, S2000magician, SDM 010, Salix alba, Salmanazar, Sgoder, Siberianmetal, Simon P, Simon J_Kissane, Sound of a desire, Stephenb, Stpasha, Stroppolo, Svick, Sympa, Tangerines, The imp, TheOriginalSoni, Timneu22, Tiroche, Tommyjs, Ushau97, Vance.naughton, Velocidex, Woohookitty, ZeroOne, Zundark, 221 anonymous edits

Kurtosis *Source:* <https://en.wikipedia.org/w/index.php?oldid=591592483> *Contributors:* 1000Faces, 212.153.190.xxx, Abocok, Adoniscik, Albmont, Andrea.granelli, Annabel, Anonymous Dissident, Antony T. Denberg, ApprentiMiami, Ashutoshroy, Audriusa, AxelBoldt, Bcdugan, BenFrantzDale, Bender235, Benwing, Blotwell, Boxplot, Chaos4.6692, Chevapdva, ChrisDok, Ciciban, Colonies Chris, Conversion script, Cruise, Czuzkatzimhut, David Haslam, DavidCBryan, Den fjättrade ankan, Docsteve.518, Donner60, DrMicro, Dragonflare82, Duoduoduo, Eric Kvaalen, Ethaniel, FelisSchrödingers, Fgnievinski, Gareth Owen, Giftlite, Glenbarnett, Graham87, GrayCalhoun, Harmeet0311, Hve, Illia Connell, JLBernstein, JavOs, Jeff DLB, Jim.belk, Jitse Niesen, Jneill, Kappa, Kowarschick, LOL, Lamro, Lebha, Leuko, Limit-theorem, Lkjhgfdsa, Lovafalk, Magnhus, MarkSweep, Me...™, Meiskam, Melcombe, Memming, Michael Hardy, Miguel, Mikeblas, Mojoworker, MrOllie, Nbarth, Nicolas Perrault III, OldakQuill, Oleg Alexandrov, PBH, Phillipkwood, Pmanderson, Policron, Pot, Quadari, Quarl, Quokly, Qwfp, Rajb245, Ramin Nakisa, Rbeech, Resprinter123, Rlendog, Rumping, S2000magician, Salix alba, Shabbychef, SimonP, Smaines, Sternthinker, Sympa, TPK, Tabako, Tanthalas39, The Gnome, The imp, Tide rolls, Tk190478, Tomi, Tpb, Twri, WikHead, Yyy, Zundark, 150 anonymous edits

Ranking *Source:* <https://en.wikipedia.org/w/index.php?oldid=589397813> *Contributors:* 16@r, A.K.A.47, Adiel, Alan Liefing, Albatross2147, AnOddName, Bentogoa, Brozen, Carrp, ChemGardener, Cherkash, Chris the speller, Colonel Tom, Curps, DAFMM, Darrel francis, David Haslam, Dr.Genius, Evand, Feco, Finnancier, Fraggel81, Franceseo Pozzi, Gaius Cornelius, Galoubet, George100, Globalphilosophy, Gr17, Gregbard, HJKeats, Hateless, Hzhing, Info hatinh, Ioksgen, Izquierdoag, Jan Dudfk, Jj2006, Jnestoros, Jhandarrington, Jpblev, JukoFF, Kiefer,Wolowitz, Kku, Kmwitko, LachlanA, Ladislav Mecir, M.nelson, Male1979, Matthewmayer, Melcombe, Mojo Hand, Mooncow, Mosmof, Msnicki, Nabeth, Nichtich, Oleg Alexandrov, Patrick, Pearle, Qwfp, R'n'B, RJFRJ, Reindeerfive, Retired username, Sigma 7, Silly rabbit, Skbkekas, SpaceFlight89, Thingg, Tony1, Twas Now, Versageek, Vertical Domination, Wooddoor, Yintan, Zzyzx11, 69 anonymous edits

Box plot *Source:* <https://en.wikipedia.org/w/index.php?oldid=591126877> *Contributors:* 127, 222fjb, 3fingeredPete, A8UDI, ALE1, AbsolutDan, Ajdlinux, Ajonlime, Alansohn, Allstarecho, AndrewHZ, Anooponnet, Ausinha, Baccyak4H, Berland, Bill william compton, BlueAmethyst, Boltor, Boxplot, BrettMontgomery, Caltas, Chen-Pan Liao, Cholmes75, ChrisGualtieri, Chrischan, Christian75, Coffee2thereums, Danharrisdnharris, David Epstein, Deljr, Dcoetzee, Den fjättrade ankan, DerHexer, Discospinster, Donner60, Dougofborg, DurDerpDurDerp, Epr123, Evercat, Fisherjs, Fnielsen, Foreurner411, Fredrik x nilsson, Fvasconcellos, G716, GVOLTT, George Brower, Giftlite, Glane23, Glrx, Gogo Dodo, GraemeL, Glane23, Glen, Glenbarnett, Hadleywickham, Headbomb, Henrygb, Hgberman, Hooperbloob, Hssghj, Hu12, IRP, Ingenue Girl, Innohead, Ion vasilieff, Iridescent, J.delaney, Jackmcbarn, Javidjamae, Jeepday, Jennavecia, Jhguch, Jim.belk, Jim1138, JoanneB, Johannes Hüsing, Johndburger, Johnjohn124, JonPeltier, Jrockley, Jwollbold, Kareekacha, Kodiologist, KuCM, Lambiam, Lourakis, Mack2, Mecanismo, Melcombe, Michael Hardy, Mpt24, MrOllie, Muel07, Mwtoews, N5ai, Nbarth, Nemo bis, Nevron, Nlu, NotTheMilkman, Notreallydavid, Noyer, Ohconfucius, Oleg Alexandrov, Oliphant, Oxymoron83, Parametrist, Pinethicket, Piotrus, Professordreamsmasher, Qwfp, RJaguar3, RandomXYZb, Res2216firestar, RexNL, Richard001, Ri, Runningonbrains, SMC, Salvio giuliano, SamaRAWR, Swoodside, Schutz, Seanstock, Skagedal, Sparklism, Startstop123, Staticshakedown, Sunroamer, Sunwards, Super-Magician, Taganov, Talgalili, The Anome, Tkirkman, Tom Duff, Tom Loughheed, Wavelength, Widr, Wikidilworth, Willking1979, Wissons, Zarat12, ZeroOne, 264 anonymous edits

Histogram *Source:* <https://en.wikipedia.org/w/index.php?oldid=591655494> *Contributors:* 160.5.82.xxx, 2D, 5 alberr square, ABF, Aastrup, AbsolutDan, Aitias, Ajraddatz, Alansohn, AlekseyP, Asterion, AugPi, Auldglory, Avenged Eightfold, Avenue, AxelBoldt, Baa, Bald Zebra, Bart133, Baticas88, BartlebytheScrivener, BenFrantzDale, BioPupil, Bkell, BlaiseEgan, Bobo192, Bogey97, Bongwarrior, Borx, Boxplot, Calvin 1998, Capt. James T. Kirk, Catalin Bogdan, Charles Matthews, Chester Markel, Chris the speller, Chris53516, ChrisGualtieri, Chunan Baka, Conversion script, Cristianrodenas, Csong12, Ctbold, Cyclopia, DVD R W, DanielPenfield, Daniele Pugliesi, DarthVader, Dcirovic, Deep316, Demus Wiesbaden, Den fjättrade ankan, Disavian, Djeeda1, Dkkicks, DmitriX, Donarreiskoffer, DrMicro, Ecov, Ehrenkater, Engahomed, Enviroboy, Epr123, EvelinaB, Evgeny, Faithlessunderboy, FallingGravity, Fatka, FayssalF, FreplySpang, Furrykef, G716, Garglebut, Gary King, Gc1mak, Gcjbkack, GeordieMcBain, GeorgeBarnick, Giftlite, Gilderien, Glane23, Glen, Glenbarnett, Hadleywickham, Headbomb, HenningThielemann, Hgberman, Hooperbloob, Hu12, Hydrogen Iodide, IGeMinX, Illia Connell, Immunize, Indon, Irishguy, Iwaterpolo, J.delaney, JRSP, James086, JamesMoose, Jamesootas, Japanese Searobin, Jerry teps, Jimmyjimjim, Jni, Johan1298, JohnBlackburne, JohnManuel, Josemiotto, Joxemai, Jusdafax, KGasso, Karol Langner, Kieran, Kjetil1001, Kku, KnightRider, Kuru, Lara bran, Lathrop, Lavaka, LeaveLeaves, Leonam, Liberio, Liftarn, Liguem, Luna Santin, MBlakley, Macrakis, Magog the Ogre, Makki98, Male1979, Manop, Martarius, Master of Puppets, Mathstat, Mattopia, Meekywiki, Melcombe, Mendaliv, Metacommet, Mhym, Michael Hardy, Mihoshi, Mogism, Moreschi, Mr coyne, Mwtoews, N5iIn, Naravorgaara, Nbarth, NeilN, Neko-chan, NerdyScienceDude, Nield, Nijdam, Noctibus, Nsaa, Ohnoitsjamie, Onionmon, P.B. Pilhet, Pathoschild, Paul August, PaulTheOctopus, Philip Trueman, Phil779, Piano non troppo, Pinethicket, Plantphoto, Podgorec, Prumpa, Pyrrhus16, Qwfp, RJaguar3, Rich Farmbrough, Richjwidr, Riley Huntley, Rjwilmsi, Robert P. O'Shea, RodC, Ronhijones, Ronz, Rufous, SD5, SE7, SFK2, Saxbryn, Swoodside, Schutz, Seraphim, Sgreewe, Shimazaki, Shoefdeath, Skizzik, Smith609, Spaz4tw, Specs112, St. Dako, Staticshakedown, Steven J. Anderson, Steven Zhang, Straif, SuperHamster, Surya21, Szalákóta, Taflykins, Tangent747, Teles, The Thing That Should Not Be, The wub, TheNewPhobia, Theswampman, Thomas Arnold, Tide rolls, Tolly4bolly, Tombrazel, Tomi, Tomleyb12, Tommy2010, Undisputedloser, Velella, Voice In The Wilderness, Vrenator, Wavelength, Webclient101, Widr, Wile E. Heresiarch, Woohookitty, Wwwwolf, Xftan, YUL89YYZ, Yamaguchi先生, Yurik, Zach1994, Zaharous, Zheric, Zhou y1777, Zondox, ZooFari, Zr40, Zvika, ماني, 477 anonymous edits

Bender235, Benwing, Berland, BernardZ, Beusson, Bevo, Bfinn, BlaiseFEgan, Blueharmony, Bobianite, Bobo192, Bomazi, Boxplot, BrendanH, Brian Crawford, Btyner, CBM, CH-stat, CardinalDan, CarlManaster, Carpentic, Cazort, Cfn011, ChangChienFu, Chris53516, Closedmouth, Coppertwig, Cosmix, Cpiral, Crash D 0T0, CremePuff222, Cretog8, Cibolt, Cybercobra, D0Kkaebi, Danilcha, DarkArcher, DavidEppstein, DavidHouse, Deimos28, Den fjättrade ankan, Denisarona, Denoir, DerHexer, Di1000, DickStartz, Dicklyon, Diegoful, Discott, Dmitronik, Doubleplusjeff, Duoduoduo, EJM86, Eaihua, EconProf86, Eli the King, Emvve, EvelinaB, Feinstein, Fleagle11, Francescapelusi, Francisbach, Friend of facts, FrozenMan, Fstonedah, Fæ, G716, GargoyleMT, Geeced, Gene Nygaard, Giftlite, Giler, Gilliam, Giogn2000, Gmelli, Gnomepirate, Goskan, Goudzovski, Gpeilon, Gprobins, Gruz, Gzkn, Hakimo99, HanPritcher, Hess88, Hike395, HughD, Hve, Hypocritical, Ilikeed, Illia Connell, Indianarhodes, Infiniti4, Ioannes Pragensis, Israel Steinmetz, J04n, JamesWatson, Jason Quinn, Jdanna, Jeremymiles, Jitse Niesen, Jmchen, Joel B. Lewis, John, JohnInDC, Jonkerz, JorisV, Joseph Solis in Australia, Julienbarlan, Jyngyr, Jérôme, KHamsun, Kgwet, Kiefer, Wolfowitz, King of Hearts, Kku, Koshier Fan, Kotsiantis, Krexer, Krishnavedala, Krubo, Kusyadi, Kvelg, LOL, Lacurus, Lealc, LiHelpa, Llinkade, Lugia2453, MZMcBride, Mangledorf, Marc K, Marcellobribeiro, Markjoseph125, Mathstat, Matthew Yeager, Mazin07, Mbhiii, Mdd, Meekohi, Melcombe, Michael Hardy, MrOllie, MrFebruary, Mwtoews, N5lin, Nbarth, NickyMcLean, Noe, Nomoskedasticity, NorsemanII, Nrcprnm2026, Nvrmm, Oleg Alexandrov, Oli Filth, Pak21, Paul August, Peepeedia, Petergans, Ph.eyes, Photonique, Pinethicket, Piotrus, Piratejosh85, Policron, Prof. Squirrel, Pruneau, Qtea, QuantumEngineer, Quickbeam, Qwfp, Qwyrxian, Qxz, RA0808, Radnagad83, RandomAct, Ravensfan5252, RenniePet, Rich Farmbrough, Rlendog, RobinH, Rocketrod1960, Rod57, Ronz, SBemper, Sabri76, Sagaciousuk, Salix alba, Samsara, Savedthat, Schwj, Sintaku, Sinxvin, SiobhanHansa, Skbkekass, Statisticsblog, Statlearn, Stephen Milborrow, Sterdeus, Stpasha, Strife911, SueHay, Sunsetsky, Talgalili, Taw, Taxman, Tbotch, Tdhaene, Tedjn, Tempodivalse, Tesi1700, The Thing That Should Not Be, Themfromspace, Thenightowl, Thomasmeeks, Tim bates, Timhowardriley, TinJack, Tolstoy the Cat, Tom.Reding, TomViza, TomyDuby, Traderlion, Tribaal, Trippingpixie, Urhixidur, Username550, Valermos, Veinor, Veryhuman, Water and Land, Wavelength, Wayward, WikHead, Wikiant, Wikid77, Wimt, Woohokitty, Woollymammoth, Wootbag, Yuanfangdelang, Zain Ebrahim111, Zfeinst, 529 anonymous edits

Path analysis (statistics) *Source:* <https://en.wikipedia.org/w/index.php?oldid=591938332> *Contributors:* Aetheling, AndreasWittenstein, BartlebytheScrivener, ChrisGualtieri, David Eppstein, Dialectic, Evidentialist, Frans2000, Geremy78, Icurite, Kbdank71, Lgallindo, MadMax, Melcombe, Michael Hardy, Mike Rosoft, Mkoval, Onco p53, Pgan002, Ph.eyes, Quantpsy, RedHouse18, Rjwilmsi, Schwjn, Tdsk, Tim bates, Vt007ken, Yungur, 23 anonymous edits

Moving average *Source:* <https://en.wikipedia.org/w/index.php?oldid=593024833> *Contributors:* A bit iffy, Adouglass, Ain92, Alex Kofman, Amatulic, Amitchaudhary, Arthana, Beetstra, Berland, Bill, Btm, Btyner, Carmitsp, Cemsbr, Chikichiki, Chipmunk, Cjs, Ckatz, CliffC, DARTH SIDIOUS 2, Daniel Quinlan, Dark Mage, DeluxNate, DerBorg, Dianna, Diannaa, Dickreuter, DragonHawk, Dreq, Econotechie, Edupedro, Ekotkie, Epr123, Esanchez7587, Euku, Falk Lieder, Falkonstar, Feco, Fenice, Foobarhoge, Gaba p, Gakmo, Gandalf61, Giftlite, Gkhanna1, Gleb, Glennimoss, GraemeL, Grin, Hdante, HenningThielemann, HimanshuJains, Hu12, Investorhenry, JLaTondre, Jamelan, JenyaTsoy, Jianingy, Jitendralovekar, Karol Langner, Kazov, Kevin Ryde, Kwertii, Lambiam, Lamro, Landroni, Lenrius, Leszek0140, Lukas227, Makeemlighter, Mandarax, Manicstreetpreacher, Martinkv, Materials scientist, Mathaddins, Maxlittle2007, Mazin07, Mehtagaaurav, Melcombe, Merosonox, Michael Hardy, MichaelZeng7, MilfordMark, Mir76, Mkacholia, Mogism, Mwtoews, Nanzen, Naught101, Naveensakhamuri, Nbarth, Netkinetic, Neurowiki, Nikai, Ninly, Paradocton, Sealcaplin, PhilKnight, Pim2009, Pisquared8, Plecch, Qwfp, R. S. Shaw, Rabarberski, Rainyepir, Ramorum, Rdhettinger, Rentier, Requestion, Richard n, Robomojo, Ron shelf, SLI, Satori Son, Scaphoph, Sid1138, Siebrun, SilaDeyo, SoledadKabocho, Soluch, Stocknet, TechAnalyster, Teorth, Thelb4, Thirsunson, Time9, TomyDuby, Tony1, Tonyho, Tooth557, Towel401, Tradematt, Tristanreid, USTUNINSEL, Utcursch, VladimirKorablin, Wa03, Wai Wai, Wavelength, Wayne Slam, WikHead, Wikid77, Wikinotforcommercialuse, Wikiolap, William Avery, Yahya Abdal-Aziz, 303, ז'אב, ז'אב, 3 anonymous edits

Student's t-test *Source:* <https://en.wikipedia.org/w/index.php?oldid=590717025> *Contributors:* A bit iffy, A3RO, ALE!, Aagtbfoua, AbsolutDan, Ace of Spades, Acroterion, Alansohn, Albmont, Alessio Facchin, Alfaisanomega, Alifirwan09, Ameliorate!, Andrew73, Andrewman327, Angelgirlatz, Anbring, Arcadian, Art Carlson, Asitgoes, Bairam, Beetstra, BenFrantzDale, Bender235, Benetto, Bentong Isles, Bgwhite, Bmj, Bobo192, Btyner, CBM, CRGreathouse, CWenger, Cato The Censor, CattleGirl, ChasingSol, Chezruli, Chowbok, Chris53516, Christian75, Cookey118, Coppertwig, DanielCD, Danstas, Darkwraith, Dashing Leech, Daughter of Mimir, Ddx, Derosatiano, Dger, Dicklyon, Digit0xin21, Dragonof, DwightKingsbury, EJM86, EPAdmirateur, Editor at Large, Edstat, Eliezg, Erab, Eribro, Eric-Wester, Erik53081, Everrettr2, Excirial, Flewis, Fraggle81, G716, Garde, Giftlite, Gjshisha, Gottinou, Gabsomepne, GregorB, Hammer Raccoon, Hamoudafg, Harput, Haruhiko Okumura, Hmjbarbosa, Hu12, Ion vasilief, Ischemia, It Is Me Here, JForget, Jeferman, JeremyA, Jonathan Hall, JorritE, Joseph.slater, Jtneill, Juan.j.l, Kevinguang, Kgwet, Lamawaale, Lastchance 000, Ldm, LizardJr8, Loodog, Lukys, Lvzon, MATThematical, MZMcBride, Majilis, MarkSweep, MarkJ789, Masarnau, MathewTownsend, Mathstat, Mdruieter, Melcombe, Michael Hardy, Midway, Mikilas, Mild Bill Hiccup, Mishnadar, Moreschi, Morioli, Movado73, Mpf3205, MrOllie, Nbaum, Naught101, Nbarth, Nbleioatts, Necromancer44, Nmg20, Noromaru, Nrcprnm2026, Oleg Alexandrov, PanTostado, Paul August, Pewwer42, Pgan002, Phuff, Picknchewz, Policron, Protonk, Pupster21, Qst, Qwfp, RT100, Radisshor, Rich Farmbrough, Richard001, Rjwilmsi, Robbyjo, Robert Ham, Robert K S, Ronz, Rubicon, Running, Ryan Vesey, SPhotographer, Sudi Sufandi, Schwjn, SeiberDoc, Sciarinæ, Swclong, Sean3000, Seglea, Selinger, Selket, ServiceableVillain, Shadowjams, Shymal, Silly rabbit, Skbkekass, Slashme, Sliver, Srleffel, Steerpike, Susko, Talgalili, Tayste, The Earwig, The1Creator, Thorwald, Tim bates, TimBock, Titanfigs, Tom Lougheed, Tom harrison, Toolnut, Tsujimasen, Turlo Lomon, Ubuntu2, Undsoweyer, Usuallylucid, Versus22, Vovcheyk, W82, Wavelength, Wikid77, Wotnow, Xenonice, Yaris678, YourEyesOnly, Yurik, Zundark, Zvika, 486 anonymous edits

Contingency table *Source:* <https://en.wikipedia.org/w/index.php?oldid=574843435> *Contributors:* A quant, Buster7, Cuvy, DVdm, Den fjättrade ankan, Ditkekov, Duke Ganote, Eleassar, Eric Kvaalen, G716, Gelingvistoj, Giftlite, Grumpfel, Hanzzoid, Headbomb, Henrygb, Ironicon, Iss246, Ixf64, Jamelan, Joxemai, Kastchei, Kent2, Kiefer, Wolfowitz, KnightRider, Laurentesfr, Martarius, Melcombe, Michael Hardy, Mormegil, Nasa-verve, Practical321, Pradtke, Qwfp, Rumping, SJP, Saxbryn, SchfiftyThree, Seglea, Talgalili, The Literate Engineer, Tirkirman, Tuxa, VivaEmilyDavies, 41 anonymous edits

Analysis of variance *Source:* <https://en.wikipedia.org/w/index.php?oldid=592329069> *Contributors:* A930913, APH, AbsolutDan, Antonov86, Ap, AppleJuggler, Araignee, Arcadian, Arthana, Baccyak4H, Bender235, Bgeelhood, Bob1960evens, Bobo192, Brighterorange, Btyner, BuddhaBubba, CBM, Chrike, Chris53516, ChrisGualtieri, Cipherswarm, Colinstu, Cxc, DaGizza, DarwinPeacock, Davidruben, Dbrødbeck, DeadEyeArrow, Den fjättrade ankan, Denisarona, Dick Belmon, Dicklyon, Dilipkumar7, Dogface, Dradamdout, Duncharris, Duoduoduo, DwightKingsbury, Dysprosia, Edstat, Erud, EverGreg, Everrettr2, Excirial, Exlibris, Eykanal, Fgnievinski, Forteblast, G716, Gak, Gamma5675, Genalipsis, Gennies, Gideon.fell, Giftlite, Gjnaasaa, Glenn netherwood, Gloridrih, Glouchs, Goskan, Grover cleveland, GroverTheGnome, Hadal, HaEb, Headbomb, Henrygb, Hersbruck, Hillel.t, Hu12, Hugo gasca aragon, I am not a dog, Illia Connell, Insanity Incarnate, Ion vasilief, Ish ishwar, Ixf64, JA(000)Davidson, JamesBWatson, Jasant, Jdgilbey, Jitse Niesen, Jj1236, Jonkerz, JosephBarillari, Jtalledo, Jtneill, Justin Mauger, K0rq, Kaihsu, Kastchei, Kembangraps, Kgwet, Kiefer, Wolfowitz, Klemen Kocjancic, Klonimus, Krishano, Kwiki, Lexor, Limnalid, Loboprof, Magioladitis, Mathstat, Maury Markowicz, Mcl, Melcombe, Memming, Mi23nen, Michael Hardy, Mohawkjohn, Mr Stephen, MrOllie, MwNRules, Mycatharsis, Nemhun, Nevillerichards, Nicke L, Nicolas Perrault III, Norndor, Novasource, Oleg Alexandrov, Orin'soren, Ostrouchov, P. Hogie, Patrick57, Peacendance, Pengortm, Peter Greenwell, Ph.eyes, Philosophicises, PhysPhD, Pinethicket, Piotrus, Pseudomonas, Quantumobserver, Qwfp, R, RDBury, RJaguar3, Ranger2006, Rich Farmbrough, Richard David Ramsey, Richard001, Richmeister, Rji, Rjwilmsi, Rocastelo, Ronz, Salix alba, Schwjn, Seglea, Silly rabbit, Skbkekass, Slowking Man, Snezy, Sofaast, Somatamoney, SsmZhang, Stefan Hartmann, Stephenb, Susko, Talgalili, Tayste, TedE, The Placebo Effect, Theking2, Tijfo98, Tlesher, Tom harrison, Tomi, TomyDuby, Travelbird, Vanderfindenna, VectorPosse, Vonkje, Vsb, Wavelength, Wesley, WhatamIdoing, Wikid, Wikid77, Woohokitty, Yintan, Yyy, 513 anonymous edits

Principal component analysis *Source:* <https://en.wikipedia.org/w/index.php?oldid=592851733> *Contributors:* A Hauptfleisch, A5, Adoniscik, Agor153, AhmedHan, Aimboy, AlanUS, Alfaisanomega, Algorithms, Amaher, Amp, AnakingAraw, AndyKali, Anthony Appleyard, Anturtle, Aproc, Archy33, BenFrantzDale, Benwing, BernardH, Bernfarr, Blackcat100, BlastOButter42, Bruce rennes, Bruguiea, Brusgadi, CRGreathouse, Carstensen, CatherineMunro, Cccddd2012, Chenhow2008, Cherkash, Chinasaur, Chire, Chosesdites, ChrisDing, ChrisGualtieri, Chuanren, Ciphers, Ck lostword, Conormct, Conormct, Crashopper, Cretchen, Crisluong, Daemun, DaveWF, Davidtrauss, Davoodshamsi, Dcoetzee, Delasck, Den fjättrade ankan, Denizstji, Denoir, Destynova, Dfbeaton, Dfrankow, Dicklyon, Discospinster, Dj.science, Dkondras, DonAByr, Dound, Dr. Crash, Dr. Submillimeter, Dront, Drusus 0, Duncanpark, Duoduoduo, Eclairs, Ed Poor, Entropeneur, Eric Kvaalen, Ericmelse, Falcorian, Fjoelskaldr, Fnielsen, Fpahl, Fran jo, Frau Holle, Frisebits, Fylbecatulous, Ga29sic, Gaba p, Gabelglesia, Gauge, GcSwRHlc, Gdm, Gene s, Germanoverlord, Giftlite, GoShow, GongYi, GromXXVII, Gufjbl, Guaka, H@r@l@d, HairyFotr, HalilYurdugul, Headlessplatter, Hechay, Helwr, HenrikMidtby, Hike395, Hilary Hou, Holon, Hovden, Hu12, Huiji, Hypersphere, Ike9898, Imarkovs, Indeterminate, Itinerant1, Ixf64, JcPriani, JPRBW, JamesXinzhiLi, Jason Davies, Jason Quinn, Jessel, Jfeckstein, Jheald, JiemingChen, Jiuguang Wang, Jmath666, Jmeppley, Jogfalls1947, JordiGH, Jorgenumata, Josve05a, Jovan, Joxemai, Jpbowen, Jtneill, Kakila, Kastchei, Kegon, Kesla, Ketiltrout, Kevin Baas, Khalid hassani, Kjetil1001, Lambiam, Larry Doolittle, Ldvbin, Lugia2453, Lumidek, Lunch, Lupin, Luwo, Lysdexia, MC10, MER-C, MaTT, Marion.cuny, Markluffel, MaxEnt, Mayur, McSly, Mcl, Mdd4696, Mdf, Mdnahas, Mehr86, Melcombe, Meredyth, Metacommet, Metsquares, Mggiganteus1, Michael Hardy, Mihai preda, Mild Bill Hiccup, Misfeldt, Mishrasknehu, MrOllie, MuellerJak, Nagarajan paramasivam, Naomi altman, Nicolasbosco, Njersyeguy, Npetteiax, NuclearWarfare, Nwstephens, Oli Filth, Ondrejspilka, OverlordQ, Oxymoron83, PCAexplorer, Pandadi, Parunach, Paum89, Peter ja shaw, Pgan002, Phil Boswell, Pijin, Pinethicket, Pjacobi, Pmg, Poline3939, Pontus, Qtea, Qwfp, R'n'B, Rcs, Rldedesma, Rich Farmbrough, RichardVeryard, Richie, Rjwilmsi, R1R1r1, Robertgreer, Robmontagna, RockMagnetist, Roymbgardner, RzR, Saforess, SamuelRiv, Sangdon Lee, SarahLZ, Shoehinger, Schewek, SchreiberBike, Seicer, Shorespirit, Shyamal, Sirhans, Sjpjantha, Skbkekass, Slashme, Slysplace, Smb1001, Smsardam, Soon Lee, Statisfactions, SteelSoul, Stevebillings, StevenDH, Susie8876, SwatiQuantie, Sylwia Ufnalska, T784303, Talgalili, Tamfang, Tayste, Tekhnofiend, Tesi1700, Thejerm, Themedie999, Thorwald, Tillman, Tmhuey, Tomer Ish Shalom, Tomi, Trovatore, User A1, User102, VasilievVV, Vecter, Vectraproject, Vincent kraeutler, Waldir, Wapcaplet, Wavelength, Whaa?, WikiMSL, Wikipedia@natividad.com, Winterstein, X7q, Yke, Zefram, Zvika, Zwilson14, ىسى, 460 anonymous edits

Diversity index *Source:* <https://en.wikipedia.org/w/index.php?oldid=584536589> *Contributors:* Amdurbin, Andycjp, AshLin, Asif Qasimov, Carabinieri, Carl Lehto, Classical geographer, Cricetus, David Blundon, Den fjättrade ankan, Dogears, DrMicro, Duncharris, Eliezg, Forbsey, Frze, GeorgeLouis, Gilliam, Hillman, Ilmari Karonen, Ilyapon, Jackhynes, Jfdarmo, Jheald, Kenadra, Kku, Lab-oratory, Lileiting, Malkinann, Melcombe, Michael Hardy, Myasuda, Nbarth, R'n'B, Richard001, Rumping, Stemonitis, Tabletop, The Anome, TimVickers, Timios, Wavelength, West.andrew.g, Widr, Wtmitchell, Xnus, Zz9296, Горушков Михаил, 39 anonymous edits

Hierarchical clustering *Source:* <https://en.wikipedia.org/w/index.php?oldid=592346761> *Contributors:* 3mta3, Ars12345, Astros4477, Chire, Cypherzero0, David Eppstein, Dmb000006, DoriSmith, Fedelebron, GTBacchus, Grscjo3, Hakkinen, Headbomb, Hike395, Ismailari, Jackieey99, Jmajf, Jose Icaza, Jy19870110, Krauss, Krishna.91, Legendre17, Mandarax, Mathstat, Mitar,

Nealmcb, NedLevine, Netzwerkerin, Piet Delpoit, Qwfp, Rjwilmsi, Robtoth1, Saitenschlager, Salih, SarahLZ, SciCompTeacher, Sgjf67, Skittleys, SleightTrickery, StuartWilsonMaui, Talgalili, Widr, 52 anonymous edits

K-means clustering *Source:* <https://en.wikipedia.org/w/index.php?oldid=591975868> *Contributors:* 0sm0sm0, 3mta3, Agor153, Alai, AlanUS, Amkilpatrick, Andkaha, AndresH, Annabel, Annusna, Ashwin, Avoide, Barabum, BenFrantzDale, BlueScreenD, CBM, Charibdis, Charles Matthews, Chire, Chrike, ChrisDing, Chrisahn, Cincoutprabu, Corvus cornix, Cronholm144, DEEJAY JPM, David Eppstein, Den fjåtræde ankan, Denshade, Duncharris, Ernepan, Fedelebrun, Fnielsen, Foma84, Foobarhoge, Gazpacho, Gfxgu, Giftlite, Golddan Gin, Greenleaf, Gringer, Gtfjbl, Hakkinen, Headbomb, Helwr, Hgkamath, Homncruse, Honkkis, Illia Connell, Illuminated, Ixf64, Jack Greenmaven, Jcallega, Jim1138, Jnothman, John of Reading, JohnBlackburne, Jonesey95, Jonsafari, Jsanchezalmeida, June8th, Killerandy, Kzafer, Leishi, Lessbread, LilHelpa, MemreCelebi, Magioladitis, Mahlon, Manyu aditya, Mark Arsten, MarkPundurs, MaterialsScientist, Mathias126, Mathstat, Mati22081979, Mauls, Mauro Bieg, Maxiittle2007, Mclnd, Melcombe, Memming, Michael Hardy, MindAfterMath, Miserlou, NedLevine, Nick Number, Ntvuok, Ostrochov, PerryTachett, Phillipe Israel, Phoolimin, Pot, Quertyus, Qwfp, Railwaycat, Ranumao, Ratiocinate, Rcalhoun, Rich Farmbrough, Ricky81682, Robert K S, SamuelRiv, Sanchom, SciCompTeacher, Simeon87, Smartcat, Soren.harward, SpuriousQ, Stimpak, Sundirac, Talgalili, Tbmurphy, Toninowiki, Turketwh, UkPaolo, Utacsecd, Watarok, Wavelength, Weston.pace, Wfolta, Wonderful597, Woolleynick, WorldsApart, Yannis1962, Zanetu, 238 anonymous edits

Matrix (mathematics) *Source:* <https://en.wikipedia.org/w/index.php?oldid=590738995> *Contributors:* 48v, ABCD, AN(Ger), ANONYMOUS COWARD0xC0DE, APerson, AbsolutDan, Adjointn, Adrian.benko, Aeriform, Aghitza, AlanUS, Alansohn, Alasdair, Aleenf1, Aleksa Lukic, Alexf, Alfred Legrand, Alucard (Dr.), AndreNatas, Andres, Angela, Anita5192, Anoko moonlight, Anonymous Dissident, Aqwis, ArnoldReinhold, Arved, Attys, Autarkaw, AxelBoldt, Azuredu, Babababoshka, Bachcell, Barnaby dawson, Bayle Shanks, Bduke, Beetstra, Bellayet, Ben R. Thomas, BenFrantzDale, Bender2k14, Bevo, Bgwhite, Bhny, BigDunc, Birat lamichhane, Bkell, Bkivdoo, Borgx, Brad7777, BrainFRZ, Brews ohare, Brianga, CRGreathouse, CactusWriter, Can't sleep, clown will eat me, Canglesea, Capricorn42, CardinalDan, CarnivorousBunny, Cerniagigante, Charles Matthews, Chinju, Chocochipmuffin, Chris the speller, ChrisMiddleton, ChrisM, Christian Matt, Codetiger, Cole the ninja, Cronholm144, Cwkmalk, CyclePat, D. Recorder, D23042304, DEMcAdams, DKqwert, DVdm, Da rulz07, Dark123, Darth Panda, David Eppstein, David Haslam, Dcljr, Ddxce, Deepmath, Derpghvdyj, DesmondSteppe, Dirac1933, Dissident, Dmcc, Doccollinni, Dogah, Donner60, DoorsAjar, Doug4, Dr.K., Dratman, Drilnoth, Dsperlich, Dtg, Duncharris, Dysprosia, EagleFan, Ejrh, El C, El Caro, Elcobbola, Epr123, Excirial, FactSpewer, FiP, Foxxwill, Francs2000, Fredrik, Freesodas, Freddie, Fresheneesz, Fritzpoll, Ftbrhygvn, Fullverse, Fusionmix, Gamer007, Gandalf61, Gazzawi, Giftlite, GirasoleDE, Gjd001, GliCh, GoonerW, HHahn, HJ Mitchell, Hajhouse, HalJor, Hans Adler, Hbent, Headbomb, Hermel, Hlevkin, IDangerMouse, INVERTED, IT2000, Igny, IkamusumeFan, It Is Me Here, ItsZippy, Ivan Štambuk, Izzedine, J.delanoy, J04n, JJ Klappack, JMK, JNW, JRM, Jacobdyer, Jagged 85, Jakob.scholbach, Jamesooders, Jao, Jarble, Jecwiki, Jezzabr, Jfheche, Jheald, Jianhui67, Jigen III, Jimmyre, Jitse Niesen, JoergenB, JohnBlackburne, JohnCD, Johnniqu, Jon Awbrey, JonMcLoone, Jordgette, Joti, Jrtayloriv, Juansempere, Kaarebrandt, Kallikanzarid, Karch, Katovatzschyn, Kevinecahill, Kiefer.Wolfowitz, KjellG, Koavf, Krishnavedala, Krun, L.O.L., La goutte de pluie, Lakeworks, Lambiam, Landroni, LeaveSleaves, Leptictidium, Levinep, LilHelpa, Livius3, LizardJr8, Lkh2099, LokiClock, Looxh, Lou Sander, Lself, LuK3, Lucaspentzlp, Lunch, LutzL, MC10, MER-C, Macrakis, Malik Shabazz, Manco Capac, Mandarax, Mandolineface, Manway, Marc van Leeuwen, MarcelB612, MarcoPotok, Marek69, Mark L MacDonald, MarkSweep, Markus Pössel, Maschen, Masnevets, Masterpiece2000, MaterialsScientist, MathInclined, MathMartin, MattTait, Maurice Carbonaro, Mazin07, Mdd, Merosonux, Mets501, Mezafoi, Mhym, Michael Devore, Michael Hardy, Michael P. Barnett, Michael Slone, Minglai, MiraiWarren, Misza13, Mlm42, Mohamed Magdy, MrOllie, Much noise, Muhandes, Muriel Gotrop, Muscularmussel, Myrvjn, Mythobeast, Nat2, NawlinWiki, Neckro, Neddyseagoon, Neparis, Nessalc, Netoholic, NevilleDNZ, Nevit, Nickmm, Nijdam, NoFlyingCars, Numbermaniac, Nurath224, Oatmealcookiemon, Obradovic Goran, Oculi, OktayD, Old Moonraker, Oleg Alexandrov, Optikos, OrthogonalFrog, OwenGage, Oxyromon83, Paolo.dL, Parages12321, Pascal.Tesson, Patrick, Paul August, Paul Murray, Pavel Vozenilek, Peldkynd, Phil888, PhotoBox, Policron, Poor Yorick, Porcher, Porges, Porphyro, Prabash.A, Prashanthns, Profvk, Proofreader77, Propower, Puffin, Quaeler, QueenCake, Quondum, Qwfp, R'n'B, R3m0t, RDBury, RIS cody, RJJFR, RPHV, Radon210, Ramin Nakisa, Rank Penguin, Rasmus Faber, Rcorcs, Reatlas, Recognizance, Remag Kee, René Vápeník, RexNL, Rgdboer, Rich Farmbrough, Richard777, Rick Norwood, Rinconsaleo, Rivertorch, Rlsheehan, Rmlison, RobHar, RobinK, Rbrurke, Rror, Rschwieb, Rucker913, Rxnt, SGBailey, Salvio giuliano, Sam need, Samuel Huang, Sander123, Saros136, Schneelocke, SchuminWeb, Scott Paeth, Senator Palpatine, SeoMac, Sfbaldbear, Shahab, Shiggity, Silly rabbit, Simone, SixWingedSeraph, Skizzik, Slawekb, Smallbones, Smallman12q, Sofia karampataki, Soupjvc, Spondoolicks, Stephen Poppitt, Steve.jaramillov, Stvertigo, Sullivan.tj, SuperHamster, Sural, Sverdrup, SwimmerOfAwesome, Swordsmankirby, Symane, Szhaidr, Slawomir Bialy, T-9000, T@nn, Tabletop, Tarquin, Tbackstr, Terry Bollinger, Tesseract. The last username left was taken, The strategy freak, The undertow, Tide rolls, TimothyRias, Toby72, Tobias Bergemann, Toghrul Talibzadeh, TomyDuby, Tosha, Tsirel, Tyler, Typofier, Tyrantbrian, U+003F, UbiquitousUK, Ugog Nizdast, UnicornTapestry, Urdutext, Urhixidur, User A.1, Username314, Urcusch, Vairoj, VasilievVV, Vonkje, W4chris, Waltpohl, Wamiq, Wanderingstan, WardenWalk, Watcharakorn, Wavelength, Wayp123, WaysToEscape, Wdrev, Webdinger, WhiteHatLurker, Wiki13, Wiml, Wolfock, Woohookitty, Wrelwser43, Wshun, Wvbalby, Xl, XJAm, Xxh1h, Yannegfrotin, Yintan, Yodalee327, Zenibus, Zeno Gantner, ZeroOne, Александр, זערו און, ২৯৩৩, ৩৯৩৩ ২৯৩৩৩৩৩৩, 841 anonymous edits

Matrix addition *Source:* <https://en.wikipedia.org/w/index.php?oldid=572737316> *Contributors:* Algebraist, AxelBoldt, BiT, Ciphers, Cruise, David Cruise, Dmcc, Ejrh, Enchanter, Erik Zachte, F=q(E+v^B), Hu12, Jan Hidders, Jeodesic, Jitse Niesen, Jérôme, Kilva, MarkSweep, Mcmark64, Mebden, NeonMerlin, Oatmealcookiemon, Octahedron80, Oleg Alexandrov, Orenburg1, PierreAbbat, Preciousena, RobHar, Rriegs, Salgueiro, Soupjvc, THEN WHO WAS PHONE?, Tesseract, Vaibhavanwal, Wshun, ZeroOne, 38 anonymous edits

Matrix multiplication *Source:* <https://en.wikipedia.org/w/index.php?oldid=593258490> *Contributors:* (Julien:), A bit iffy, A little insignificant, APerson, Acabit, Achurk, Acolombi, Adrian 1001, Alephhaz, AlexG, Ambhrani, Anonash, Aqwis, Arcfrk, Arthur Rubin, Arvindn, AugPi, AxelBoldt, BenFrantzDale, Bender2k14, Bentogoa, Bgwhite, Bkell, Boud, Brian Randell, Bryan Derksen, Calbaer, Calmer Waters, CanaDanijl, Churnett, Chewings72, Chris Q, Christos Boutsidis, Citrus538, Cloudmichael, Coder0xf2, Coffee2theorems, Copyedit0r42, Countchoc, Damian Yerrick, Damirgraffiti, Dandin1, Dbroadwell, Dcoetzee, Dekart, Denisarona, DennyColt, Dominus, Donner60, Dooblebop, Doshell, Dratman, Ejrh, ElizabethRogers, Erik Zachte, Ethically Yours, F=q(E+v^B), Fbahr, Felmira, Forderud, Fresheneesz, FrozenUmbrella, Gandalf61, Gauge, Ged.R, Giftlite, Gryllida, HmSSolent, Haham hanuka, Happy-melon, Harris000, Hassé Weyl, Headbomb, HenningThielemann, Hermel, Hmonroe, Hoyvin-Mayvin, InoShiro, InverseHypercube, Irina1222, JakeVortex, Jakob.scholbach, Jarble, Jitse Niesen, Jivee Balu, JohnBlackburne, JohnMathTeacher, Jon Awbrey, Joshua, Jérôme, K.menin, Kaluzman, KelySYC, Kevin Baas, Kmel, Koertefa, Kri, Kvg, Lakeworks, Lambiam, Liao, LkNsgnth, LouScheffer, Man It's So Loud In Here, Marc Venot, Marc van Leeuwen, Mariusepicurean, Maschen, Mate2code, MathMartin, MattTait, Max Schwarz, Mdd4696, Melchoir, Mellum, Michael Hardy, Michael Slone, Midoreigh, Miqonranger03, Miy, Mxhan3189, Mononomic, MrOllie, Msalimi1222, MuDavid, Inverse, NellieBly, NeonMerlin, Neparis, Ngvrnd, Nijdam, Nikola Smolenski, Nobar, Nsk92, Olathe, Oleg Alexandrov, Oli Filth, Optikos, OrangeKyo, PMLawrence, PV=nRT, Paolo.dL, Parerga, Patrick, Paul August, Paul D. Anderson, Paul Wolneykien, Pfnlo, Psychlohexane, Pt, Quartl, Quondum, Quertyus, R.e.b., RDBury, Radius, Ratiocinate, RexNL, Rhsimard, Risk one, Robertwb, Rockn-Roll, Rpspeck, Rschwieb, Running, Salih, Sandeep.murthy, Schneelocke, Scientific29, Shyubchef, Shahab, Skaraoko, Slawekb, Snorkelman, Sonicyouth86, Sr3d, Ssola, Stephane.magneat, Sterrys, Lou Sander, SwickProgramming, Svick, Swift1337, SyntaxError55, Slawomir Bialy, Tarquin, Terry Bollinger, The Fish, The Thing That Should Not Be, TheGeomaster, Thenub314, Throwaway85, Tide rolls, Tobias Bergemann, Tyrantbrian, Umofomia, Uri-Levy, VAXHeadroom, Vgmgddg, Vincinsonfire, Virginia-American, Wavelength, Widr, Will Thompson, WizMystery, Wshun, Xypron, Zazpot, Zdmitrak, Zero0000, Zhangleisk, Zmaboros, Zorakoid, 350 anonymous edits

Transpose *Source:* <https://en.wikipedia.org/w/index.php?oldid=591637045> *Contributors:* Adam Zivner, Amakuha, Anaxial, Andres, Army1987, AxelBoldt, BenFrantzDale, Bogdanno, Calle, Cantus, Catslash, CenturionZ 1, Cesiumfrog, Chewings72, ChrisGualtieri, Cwkmalk, DHN, Dysprosia, EconoPhysicist, El C, Filip13041982, Fintor, Frietjes, FvdP, Gail, Gandalf61, Geometry guy, Giftlite, Ideyal, J.delanoy, J04n, JEBrown87544, JabberWok, Jasonyo, Javalonok, Jingsiaoting, Jitse Niesen, JoeDub, Joriki, Jxramos, Kausinghatak, Kilui, KlappCK, Kwamikagami, Landroni, Larry V, Laurentius, Lethe, LokiClock, LucasVB, Lunch, Lyhana8, LucaTait, MattTait, Mestrother, Metasquares, Mets501, Michael Hardy, Mozk6, Nm26586, Nbarth, Neparis, Noideta, O18, Octahedron80, Oleg Alexandrov, Oli Filth, Palanq, Pomona17, Pomte, Pred, Quondum, Quoxplusone, Raffamaiden, Rludlow, Robpbrain, Ronhjones, SeventyThree, Slawekb, Smjg, Stangaa, Stevenj, TakuyaMurata, Talgalili, Tarquin, TeH nOmInAtOr, Template namespace initialisation script, Tobias Bergemann, Tokek, Tom Toyosaki, Vkpdl11, Wavelength, WissensDürster, Wwoods, 90 anonymous edits

Determinant *Source:* <https://en.wikipedia.org/w/index.php?oldid=592998743> *Contributors:* 01001, 165.123.179.xxx, A-asadi, A. B., AbsolutDan, Adam4445, AdamP, Ae77, Ahoerstemeier, AlanUS, Alex Sabaka, Alexander Chervov, Alexandre Duret-Lutz, Alexandre Martins, Alexey Muranov, Algebraist, Alison, Alkarex, Alksub, Anakata, Andres, Anonymous Dissident, Anskas, Ardonik, ArnoldReinhold, Arved, Asmeurer, AugPi, AxelBoldt, BPets, Balagen, Barking Mad142, BenFrantzDale, Benchartlett, Bender2k14, Benwing, Betacommand, Big Jim Fae Scotland, BjornPoonen, Bkonrad, BrianOfRugby, Bryan Derksen, Burn, CBM, CRGreathouse, Campuzano85, Camrn86, Carbonrodney, Catfive, Cbogart2, Ccandan, Cesarth, Charles Matthews, Chester Markel, Chewings72, Chocochipmuffin, Christopher Parham, Cjkstephenson, Closedmouth, Cmnhf5, Cobi, Coffee2theorems, Colombomatia89, Connely, Conversion script, Correnos, Cowanae, Crossover, Cronholm144, Crystal whacker, Cthulhu.mythos, Cutelyaware, Cuzkatzimhut, Cwkmalk, Danaman5, Dantestyrael, Dark Formal, Datahaki, Dcoetzee, Decora, Delirium, Delatadron, Demize, Dmbrown00, Dmcc, Doctormatt, DonAbyrd, Donner60, Dysprosia, E290341, EconoPhysicist, Edwardrf, Elphing, Eniagrom, Entropeneur, Epr123, EtudiantEco, Euphrat1508, Everest05, Execute Outcomes, Ffatosajavazii, Fredrik, Froppuf, Gabriel10yf, Gauge, Gejkeji, Gene Ward Smith, Gershwinnr, Giftlite, Graham87, GrewalWiki, Greynose, Guiltyspark, Gwernol, Hangifresh, Headbomb, Heili.brenna, Henkvd, HenningThielemann, Hkmcsczz, Hlevkin, Ian13, Icaims, Icktoofay, Ijpluido, Improbable keeler, Ino5thor, Istcol, Itai, JC Chu, JJ Harrison, JackSchmidt, Jackzhp, Jagged 85, Jakob.scholbach, Jasonevans, Jeff G., Jemeibus, Jerry, Jersey Devil, Jewbacca, Jheald, Jim.belk, Jitse Niesen, Joeej, Jogers, Johnniqu, Jondaman21, Jordgette, Joriki, Josp-mathilde, Josteinaj, Jrgetsin, Jshen6, Juansempere, Justin W Smith, Kaarebrandt, Kallikanzarid, Kaspar.jan, Kd345205, Khabog, Kingpin13, Kmhkmh, Kokin, Kstueve, Kunal Bhalla, Kurykh, Kwantus, L.Ancienne, L.O.L., Lagelspiel, Lambiam, Lavaka, Leakejee, Lethe, Lhf, Lightmouse, LilHelpa, Logapragasan, Luiscardona89, MackSalmon, Marc van Leeuwen, Marek69, Marozols, MartinOtter, Maschen, MaterialsScientist, MathMartin, McKay, Mcconnell13, Mestrother, Mdnahas, Merge, Mets501, Michael Hardy, Michael P. Barnett, Michael Slone, Mikael Häggström, Mild Bill Hiccup, Misza13, Mxhan3189, Mmxx, Mobiusstheof, Mrsaad31, Msa1Iusec, MuDavid, Myshka spasayet Iva, N3vln, Nachiketvartak, Nat2, Neilen Marais, Nekura, Netdragon, Nethgurb, Netrap, Nickj, Nicolae Coman, Nistra, Nsaa, Numbo3, Obradovic Goran, Octahedron80, Oleg Alexandrov, Oli Filth, Paolo.dL., Patamia, Patrick, Paul August, Pedrose, Pensador82, Personman, PhysPhD, Pigei, Priitliivak, Protonk, Pt, Quadell, Quadrance, Quantling, Quondum, R.e.b., RDBury, RIBEYE special, Rayray28, Rbb 1181, Rdmabry, Recentchanges, Reinyday, RekishiEJ, René Vápeník, RexNL, Rgdboer, Rich Farmbrough, Robinh, Rocchini, Roentgenium111, Rogper, Rpchase, Rpytle731, Rumbelthunder, SUL, Sabri76, Salgueiro, Samiswicked, Sandro.bosio, Sangwine, Sayahoy, SchreiberBike, Sebhoffer, Shai-kun, Shreevatsa, Siener, Simon Sang, SkyWalker, Slady, Smitherens, Snoves, Spartan S58, Spireguy, Spoon!, Ssd, Stdzia, Stefano85, Stevenj, StradivariusTV, Sun Creator, Supreme fascist, Swerdraneb, SwordSmurf, Slawomir Bialy, T8191, Tarif Aziz, Tarquin, Tau, TedPavlic, Tegla, Tekhnofind, Tgr, The Thing That Should Not Be, TheEternalVortex, TheIncredibleEddieOompaLoompa, Thehelpfulone, Thenub314, Timberframe, Tobias Bergemann, Tolly4bolly, TomViza, Tosha, Trashbird1240, TreyGreer62, Trifon Triantafillidis, Trivialefault, Trompedo, Truthnlove, Ulisse0, Unbitwise, Urdutext, Vanka5, Vibhor1997, Vincent Semeria, Wael Ellithy,

Wavelength, Westwood0.137, Wik, Wirawan0, Wolfrock, Woscalfrench, Wshun, Xiaos, Zaslav, Zutuz, Zzedar, ^demon, 487 anonymous edits

Minor (linear algebra) *Source:* <https://en.wikipedia.org/w/index.php?oldid=585593096> *Contributors:* Addshore, Archimerged, AugPi, AxelBoldt, BD2412, Bohumir Zamecnik, Campuzano85, Charles Matthews, Dysprosia, Eroblar, Forderud, Fresheneesz, Gaius Cornelius, Gauge, Giftlite, Knakts, Mark L MacDonald, MarkSweep, Mynamesrlyweirdhawhaw, Nbarth, Neparis, Nicinic, Nihiltres, Paolo.d.L., Pavlovič, Pdeq, Pmdboi, PoomK, Quondum, Robertas.Vilkas, Simogasp, Spinningspark, Tarquin, TeH nOmInAtOr, TedPavlic, WojciechSwiderski, 金野裕希, 37 anonymous edits

Adjugate matrix *Source:* <https://en.wikipedia.org/w/index.php?oldid=584626767> *Contributors:* A Thousand Doors, Albert0168, Ali 24789, Alvar1007, Amtiss, Atomician, AugPi, AxelBoldt, Bohumir Zamecnik, Booyabazooka, Chow515, Charles Matthews, DBigXray, DFTDER, Divyatyam, Dysprosia, Ed Gibbon, El C, Elphion, Fernandopabon, GeorgeOne, Giftlite, Halo2, Holek, JabberWok, Jhschenker, Jiejunkong, KSmrq, Klemen Kocjancic, Koliokolio, Kyros1, Lunch, Marc van Leeuwen, MarchHare, MaterialsScientist, Menthoolium, Merge, Michael Hardy, Michael Slone, Mkill, MuDavid, Neparis, Nixphoeni, Oleg Alexandrov, Paolo.d.L., Peskydan, Ptheoch, Quantling, Quondum, René Vápeník, Rht1369, Rossengers, Ryms84, SUL, Salgueiro, Scentoni, Schneelocke, Shashank2303, StijnDeVuyt, Tarquin, Teika kazura, TheJJJunk, TooMuchMath, Torquemada0, Toufiq Arafat, Trompedo, Ulisse0, Vladkpono, 139 anonymous edits

Invertible matrix *Source:* <https://en.wikipedia.org/w/index.php?oldid=590639281> *Contributors:* A3RO, ABoerma, Aaron Hill 2, Abdelrahman Mohamed Sayed, Adam Nohejl, Admdikramr, Aeris-chan, Anarnet, Andres, Antares5245, Arauzo, Arbitrarily0, Austinmohr, AxelBoldt, Ayhanbilgin, Basten, BayesianLogik, Ben pcc, BenFrantzDale, Breqwas, CBM, Cairomax, Calliopejen, Calwiki, Catskineater, Charles Matthews, ChrisGualtieri, Cmansley, DHN, DVdm, Darij, Davidhorman, Deoetzee, Deeptrivia, Dino, Discospinster, Duoduoduo, Dzhim, EconoPhysicist, EfiJalkowski, EmiIJ, Epistemical, Eraserhead1, Error792, Falcor84, Fgdorais, Fgnievinski, Frederic Y Bois, Fredrik, Furrykef, G37x8004uc, Giftlite, Godzatswing, GregorB, Grinevitski, Gsookolov, Heiko1980, Hu12, InvariantRob, InverseHypercube, Isopropyl, JEBrown87544, Jacj, Jarble, Jashar, Jitse Niesen, JohnBlackburne, JohnMathTeacher, JordiGH, K.menin, KHamsun, Kaarebrandt, Kanie, Kiefer.Wolfowitz, Kingpin13, Kkddkkdd, Kostmo, Kri, LOL, Landroni, Larrybaxter, Latanius, Leapfrog314, LOKIClock, Lukax, ML, Marc van Leeuwen, Mardetanha, Mark L MacDonald, MarkSweep, Markus Schmaus, Martin Kraus, MathMartin, Matikapoika, MattTait, Mecanismo, Miaow Miaow, Michael Hardy, Michael Weitzel, NYKevin, Neparis, Netkinetic, Nixphoeni, Obradovic Goran, Ojigiri, Oleg Alexandrov, Oli Filth, Oliphanta, PMLawrence, Penitence, PhotoBox, Policron, Poor Yorick, Protector, Quantling, Raffamaiden, Ralivingston, Rhuso, Rudlow, Robinh, Scalar F, Scching, Sdfstc, Selfworm, Shay Guy, Silly rabbit, Simamura, Sir Edward V, Slaunger, Squizzz, Stangaa, StradivariusTV, Styrofoam1994, Svick, Slawomir Bialy, THEN WHO WAS PHONE?, TedPavlic, TeleComNasSprVen, Teque5, Thecheesykid, Tobias Bergemann, TomViza, Trompedo, Turidoth, Ulisse0, Urdutext, Vovchyyk, Vrenator, Wavelength, Wshun, X7q, Zath42, Zhangmoon618, Zvika, 虞海, 271 anonymous edits

Eigenvalues and eigenvectors *Source:* <https://en.wikipedia.org/w/index.php?oldid=591739790> *Contributors:* 123Mike456Winston789, 336, 83d40m, A civilian, A930913, AManWithNoPlan, ANONYMOUS COWARD0xCODE, Aacool, Aarongeller, Abstracte, Acabashi, Adam78, AdamSmithee, Adpadu, Adpete, Agthorr, Ahmad.tachyon, AhmedHan, Ajto8, Alainr345, AlanUS, Alansohn, Albmont, AlexCornejo, Aliekens, Almit39, Alouslybum, Amberrock, Amgtech, Ancheta Wis, Anoko moonlight, Bacyyak4H, Bdegfcumbbfv, BenFrantzDale, Bender2k14, Benzi455, Bevo, Bgwhite, BillC, Billymac00, Bissinger, Bjelleklang, Bkittel, Blanerhoads, Blotwell, Bobguy7, Bochev, Booyabazooka, Boris Alexeev, Bovlb, Boyzindahoos, Brad7777, Brian0918, Buster79, Butwhatdoiknow, CRGreathouse, Capagot, Chichui, Chire, ChrisGualtieri, Chrisbaird.ma, Christian75, CinchBug, Circus, Cleared as filed, Colonel angel, Complexica, Connelly, Conscious, Cosmikz, Cotterr2, Crasshopper, Crowsnest, Crunchy Numbers, Crust, Curtdbz, Cwkmal, Cyde, D1ma5ad, DanBri, Danger, David Binner, David Eppstein, DavidFHoughton, Davigoli, Daytona2, Deoetzee, Delirium, Denwid, Dependent Variable, Deryck Chan, Dfsison, Dhollm, Dima373, Dirac66, Dmazin, Dmharvey, Dmn, Doleszki, Dotancohen, Dr. Nobody, DragonflySixtyseven, Dratman, Dwwaddell, Dysprosia, EFZR090440, EconoPhysicist, Edinborgarstefan, Editor at Large, Edsanville, Efikman, Ekwy, Elliottt, Eric Forste, Erxnmedia, Etr52, Fgnievinski, Finell, Fintor, Flyer22, Foobarnix, Forbes72, Forderud, Fortd33, FrankFlanagan, FreplySpang, Fresheneesz, Frizzil, Gaius Cornelius, Gak, Gandalf61, Gareth Owen, Gbnogkfs, Gdormer, Giftlite, Gimbeline, Giulioopp, Grubber, Gibanta, Guardian of Light, Gunderburg, GunnerJr, Gwideman, H1voltage, Haeleth, Hairy Dude, Hankel operator, HappyCamper, Haseldon, HcorEric X, Headbomb, Hede2000, Hetar, Hiiiiiiiiiiiiiiiiiiii, Hitman012, Hongooi, Humanenrg, Hunter.moseley, Hydrogravity, Iainscott, Ichakrab, Igny, Incnis Mrsi, InvictaHOG, Iridescent, Itsmine, J. Finkelstein, JMK, JPD, JYUuyang, JaGa, JabberWok, JahJah, Jakarr, Jakew, Jakob.scholbach, Jamesjhs, Javalenok, Jayden54, Jearroll, Jeff560, JeffAEdmonds, JeffieAlex, Jefromi, Jheald, JinPan, Jitse Niesen, Joel31, JohnBlackburne, Johnpacklambert, Jok2000, Jorge Stolfi, JosephCatrambone, Josh Cherry, Josp-mathilde, Jtwdog, JuPitEer, Justin W Smith, Jérôme, KHamsun, Kanie, Kanonkas, Kappa, Katefan0, Kausikhatak, Kedwar5, Keenan Pepper, Kevinj04, Kiefer.Wolfowitz, Kier07, Kimbly, Kjoonlee, Kmote, Kri, Krucraft, Kungfuadam, LBehounek, LOL, LaQuilla, Lacatosis, Lalahuma, Landroni, Lantonov, Laplacian, Larryisgood, Lepfermandes, LkNsngh, LOKIClock, Lone Isle, LordViD, Lowellian, LucasVB, Luk, Luna Santin, Luolimao, Lzyvzl, M4ry73, Madanon, Magister Mathematicae, Male1979, Mandolinface, Manixer, MarSch, Marc van Leeuwen, MarcelB612, Mark L MacDonald, Markus Schmaus, MartCMdeJong, Martyulrich, Marudubshinki, Matthewmoneck, Matrix, Maurice Carbonaro, Maziar.irani, McKay, Mcstrother, Mct mht, Mebden, MedicineMan555, Menthoolium, Metaeducation, Michael Hardy, Michael Slone, MichaelBillington, Mikhail Ryazanov, Moala, Moonraker12, Moriori, Ms2ger, Muhandes, Mushin, Mxipp, Myasuda, Napalm Llama, Natrij, NatusRoma, Nbarth, NewEconomist, Nichalp, Nick Number, Nickshanks, Nigellwh, Nihonjoe, NinjaDreams, NormDor, Not a dog, Ntjohn, ObsessiveMathsFreak, OIEnglish, Oleg Alexandrov, Oli Filth, Orizon, Ouzel Ring, Oxygene123, Pablo.e, Paolo.d.L., Patrick, Patrick0Moran, Pedant, Phyrexicaid, Piano non troppo, Plastikspork, Plrk, Pmanderson, Policron, Porejide, Protonk, Pscrape, Pt, Pushkar3, Qiangshiweiguan, Quondum, R'n'B, RJFJR, RProgrammer, Raffamaiden, Rajah, Randomblue, Rbanzai, Rchandan, Reaverdrop, Red Act, Reddevyl, Repliedthemockturtle, Restu20, Rexas, Rgdboer, Rich257, Rick Norwood, Riteshsood, Rjwilmsi, Rlupsa, Rmbyoung, Rohitphy, Rspcer, Rubybrian, Rushiagr, Ruslan Sharipov, Ruud Koot, Rxnt, Saburr, Safalra, Saketh, Salix alba, Sameerkale, Sanchom, Schismata, Schutz, Scott Ritchie, Sebastian Klein, Severon, Shai-kun, Sherif helmy, Shishir0610, Shizhao, Shreevatsa, Silly rabbit, Simon12, Skakkle, Skittleys, Slawekb, Smit-Meister, Somesh, Soultaco, Spikey, Srich32977, Sschongster, Ste4k, Stephen Poppitt, Stevelinton, StevenJohnston, Stevenj, Stevertigo, StradivariusTV, Sunray, Szabolcs Nagy, Slawomir Bialy, TDogg310, TVilkasalo, Tabletop, Tac-Tics, Taco325i, Tarquin, Tator2, Tatpong, TeH nOmInAtOr, TeaDrinker, TerraNova, whatcanidotmakethisnottoosimilarosomeothername, Tesi1700, The Duke of Waltham, The suffocated, TheRealInsomnius, Thecheesykid, Theneb314, Thorfinn, Timeroot, Timhooey, Timrollpickering, Tiny green, Titoxd, Tkuvho, Tobias Bergemann, Tomo, TomyDuby, TreyGreer62, Trifon Triantafillidis, Ttennebkram, TypoBoy, Tyraios, Urdutext, Urtis, User A1, Varuna, Vaughan Pratt, Vb, Veganaxos, Waldir, Wavelength, Wayp123, Wayward, WhiteHatLurker, Wikid77, Williampoetra, Winston Trechane, Woohookitty, Wootery, Xantharius, Xelnx, Xnn, Yahya Abdal-Aziz, YoshigeV, Yurik, Zapurva, Zaslav, ZeroOne, Zinnmann, Zylinder, ٧٩١٣٧٩١, 827 anonymous edits

System of linear equations *Source:* <https://en.wikipedia.org/w/index.php?oldid=592807301> *Contributors:* 3rdiw, Alikhtarov, Amahoney, Andres, Anita5192, Anwar saadat, Arctic Kangaroo, Arthana, Arthur Rubin, AxelBoldt, Azuredu, Brainfck, Braveorca, Caesura, Calle, Caribbean H.Q., Charles Matthews, Chris the speller, CinchBug, Constructive editor, Crazycomputers, Cybercobra, D.Lazard, DHN, Daniel Brockman, Danroa, Day and Nite, Denisarona, Dgw, Dgwarwick, Dhatfield, Discospinster, Dobermanji, DotKuro, Driski555, Duoduoduo, Dysprosia, Erkan Yilmaz, Everyking, Frankie0607, Freddie, Gesslein, Giftlite, Giro720, GorillaWarfare, Hao2lian, Hgkamath, Hydrogen Iodide, II MusLiM HyBRiD II, IgorCarron, Inframaut, InverseHypercube, Ivan Štambuk, J.delanoy, JPM-GR, Jauhienij, Jeff3000, Jim.belk, JinJian, Jitse Niesen, John of Reading, Joriki, K.menin, KGasso, KSmrq, KYN, Kablammo, Kaspar.jan, Khalid Mahmood, Kku, LOL, Lambiam, Larry Doolittle, Makalrfekt, Mark L MacDonald, MartinOtter, Mathemaduenn, Mattoothman, Mets501, Mhss, Michael Slone, Newyorxico, Nk, Noyder, Nvrnmd, Obradovic Goran, Ooz dot ie, Ott0, Paul August, Pieter Kuiper, Plastikspork, Pratyush Sarkar, Quadedl, RProgrammer, Ranveig, Rchrd, René Vápeník, Richard B. Frost, SahilK7654, Salix alba, Sam Hoocevar, Semifinalist, Slawekb, Spiel496, Steve.jaramillov, Stigmatella aurantiaca, Svick, Tarquin, Tavernenses, Thingg, Tiddly Tom, Tim32, ToLLa, Tommy2010, Torokun, Trifon Triantafillidis, Twri, Universalss, Urdutext, Vanakariss, Vegard, Wavelength, Webclient101, Wikijens, Willtron, Wknight94, Wmasterj, Woohookitty, Wshun, Yerpo, Zchenyu, Zingus, Zserghei, Zzyzx11, ^demon, 217 גרשון, 2 anonymous edits

Image Sources, Licenses and Contributors

Image:Comparison mean median mode.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Comparison_mean_median_mode.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Cmglee

File:Comparison mean median mode.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Comparison_mean_median_mode.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Cmglee

file:Scaled chi squared.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Scaled_chi_squared.svg *License:* Creative Commons Zero *Contributors:* User:Jheald

file:Scaled chi squared cdf.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Scaled_chi_squared_cdf.svg *License:* Creative Commons Zero *Contributors:* User:Jheald

File:standard deviation diagram.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Standard_deviation_diagram.svg *License:* Creative Commons Attribution 2.5 *Contributors:* Mwtoews

File:cumulativeSD.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:CumulativeSD.svg> *License:* Public Domain *Contributors:* Normal_Distribution_CDF.svg: Inductiveload derivative work: Wolfkeeper (talk)

File:Comparison standard deviations.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Comparison_standard_deviations.svg *License:* Public Domain *Contributors:* JRBrown

Image:Nuvola apps kchart.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Nuvola_apps_kchart.svg *License:* GNU Lesser General Public License *Contributors:* en:David Vignoni, User:Stannered

Image:standard deviation diagram.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Standard_deviation_diagram.svg *License:* Creative Commons Attribution 2.5 *Contributors:* Mwtoews

Image:SkewedDistribution.png *Source:* <https://en.wikipedia.org/w/index.php?title=File:SkewedDistribution.png> *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* Audrius Meskauskas

Image:Negative and positive skew diagrams (English).svg *Source:* [https://en.wikipedia.org/w/index.php?title=File:Negative_and_positive_skew_diagrams_\(English\).svg](https://en.wikipedia.org/w/index.php?title=File:Negative_and_positive_skew_diagrams_(English).svg) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Rodolfo Hermans (Godot) at en.wikipedia.

Image:KurtosisChanges.png *Source:* <https://en.wikipedia.org/w/index.php?title=File:KurtosisChanges.png> *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* Audrius Meskauskas

File:1909_US_Penny.jpg *Source:* https://en.wikipedia.org/w/index.php?title=File:1909_US_Penny.jpg *License:* Public Domain *Contributors:* Nicolas Perrault III

Image:Pearson type VII distribution PDF.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Pearson_type_VII_distribution_PDF.png *License:* GNU General Public License *Contributors:* Jochen Burghardt, MarkSweep, 1 anonymous edits

Image:Pearson type VII distribution log-PDF.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Pearson_type_VII_distribution_log-PDF.png *License:* Public Domain *Contributors:* MarkSweep, 2 anonymous edits

Image:Standard symmetric pdfs.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Standard_symmetric_pdfs.png *License:* GNU General Public License *Contributors:* User:MarkSweep

Image:Standard symmetric pdfs logscale.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Standard_symmetric_pdfs_logscale.png *License:* Public Domain *Contributors:* User:MarkSweep

Image:Michelsonmorley-boxplot.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Michelsonmorley-boxplot.svg> *License:* Public Domain *Contributors:* User:Mwtoews, User:Schutz

File:Box-Plot mit Min-Max Abstand.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Box-Plot_mit_Min-Max_Abstand.png *License:* unknown *Contributors:* Schlurcher

File:Box-Plot mit Interquartilsabstand.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Box-Plot_mit_Interquartilsabstand.png *License:* unknown *Contributors:* Schlurcher

File:Fourboxplots.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Fourboxplots.svg> *License:* Creative Commons Zero *Contributors:* User:BrettMontgomery

Image:Boxplot vs PDF.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Boxplot_vs_PDF.svg *License:* Creative Commons Attribution-Sharealike 2.5 *Contributors:* Original uploader was Jhguch at en.wikipedia; Derivative work: Chen-Pan Liao (talk).

Image:Histogram of arrivals per minute.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Histogram_of_arrivals_per_minute.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* DanielPenfield

File:Black cherry tree histogram.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Black_cherry_tree_histogram.svg *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* Mwtoews, 2 anonymous edits

File:Travel time histogram total n Stata.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Travel_time_histogram_total_n_Stata.png *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* -

File:Travel time histogram total 1 Stata.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Travel_time_histogram_total_1_Stata.png *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Qwfp (talk)

File:Cumulative vs normal histogram.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Cumulative_vs_normal_histogram.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Kieran

File:normal exponential qq.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Normal_exponential_qq.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Skbkek

File:normal normal qq.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Normal_normal_qq.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Skbkek

File:weibull qq.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Weibull_qq.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Skbkek

File:ohio temps qq.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Ohio_temps_qq.svg *License:* Creative Commons Attribution 3.0 *Contributors:* Skbkek

File:State Route 20.png *Source:* https://en.wikipedia.org/w/index.php?title=File:State_Route_20.png *License:* Public Domain *Contributors:* Walter Siegmund

File:PD-icon.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:PD-icon.svg> *License:* Public Domain *Contributors:* Alex.muller, Anomie, Anonymous Dissident, CBM, MBisanz, PBS, Quadell, Rocket000, Strangerer, Timotheus Canens, 1 anonymous edits

Image:Flammability diagram methane.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Flammability_diagram_methane.png *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Power.corrupts

Image:Ag-Au-Cu-colours-english.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ag-Au-Cu-colours-english.svg> *License:* Creative Commons Attribution-Share Alike *Contributors:* Original image: Metallos

image:HowToCalculatePercentCompositions Altitude Method.gif *Source:* https://en.wikipedia.org/w/index.php?title=File:HowToCalculatePercentCompositions_Altitude_Method.gif *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Clex.Alark

image:HowToCalculate%Compositions Intersection Method.gif *Source:* https://en.wikipedia.org/w/index.php?title=File:HowToCalculate%Compositions_Intersection_Method.gif *License:* unknown *Contributors:* -

image:ternary.example.1.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ternary.example.1.svg> *License:* Public Domain *Contributors:* cflm (talk)

image:ternary.example.axis.1.jpg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ternary.example.axis.1.jpg> *License:* GNU Free Documentation License *Contributors:* Statistics

image:ternary.example.axis.2.jpg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ternary.example.axis.2.jpg> *License:* GNU Free Documentation License *Contributors:* Statistics

image:ternary.example.axis.3.jpg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ternary.example.axis.3.jpg> *License:* GNU Free Documentation License *Contributors:* Statistics

image:Ternary plot 1.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Ternary_plot_1.png *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Power.corrupts

image:Ternary plot 2 (reverse axis).png *Source:* [https://en.wikipedia.org/w/index.php?title=File:Ternary_plot_2_\(reverse_axis\).png](https://en.wikipedia.org/w/index.php?title=File:Ternary_plot_2_(reverse_axis).png) *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Power.corrupts

image:Ternary plot.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Ternary_plot.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Random7

File:ternary plot visualisation.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Ternary_plot_visualisation.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Cmglee

image:ternary.example.plot.1.jpg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ternary.example.plot.1.jpg> *License:* GNU Free Documentation License *Contributors:* Statistics

image:ternary.example.plot.2.jpg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ternary.example.plot.2.jpg> *License:* GNU Free Documentation License *Contributors:* Statistics

image:ternary.example.plot.3.jpg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ternary.example.plot.3.jpg> *License:* GNU Free Documentation License *Contributors:* Statistics

image:ternary.example.plot.4.jpg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ternary.example.plot.4.jpg> *License:* GNU Free Documentation License *Contributors:* Statistics

image:ternary.example.plot.5.jpg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Ternary.example.plot.5.jpg> *License:* GNU Free Documentation License *Contributors:* Statistics

File:Normal Distribution PDF.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Normal_Distribution_PDF.svg *License:* Public Domain *Contributors:* Inductiveload

File:Normal Distribution CDF.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Normal_Distribution_CDF.svg *License:* Public Domain *Contributors:* Inductiveload

File:OEISicon light.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:OEISicon_light.svg *License:* Public Domain *Contributors:* Billinghurst, Mate2code, Senator2029

File:De moivre-laplace.gif *Source:* https://en.wikipedia.org/w/index.php?title=File:De_moivre-laplace.gif *License:* Public Domain *Contributors:* Spasha

File:Dice sum central limit theorem.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Dice_sum_central_limit_theorem.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Cmglee

File:QHarmonicOscillator.png *Source:* <https://en.wikipedia.org/w/index.php?title=File:QHarmonicOscillator.png> *License:* GNU Free Documentation License *Contributors:* en:User:FlorianMarquardt

File:Fisher iris versicolor sepalwidth.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Fisher_iris_versicolor_sepalwidth.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* en:User:Qwfp (original); Pbroks13 (talk) (redraw)

File:FitNormDistr.tif *Source:* <https://en.wikipedia.org/w/index.php?title=File:FitNormDistr.tif> *License:* Public Domain *Contributors:* Buenas días

File:Planche de Galton.jpg *Source:* https://en.wikipedia.org/w/index.php?title=File:Planche_de_Galton.jpg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Antoine Taveneaux

File:Carl Friedrich Gauss.jpg *Source:* https://en.wikipedia.org/w/index.php?title=File:Carl_Friedrich_Gauss.jpg *License:* Public Domain *Contributors:* Gottlieb BiermannA, Wittmann (photo)

File:Pierre-Simon Laplace.jpg *Source:* https://en.wikipedia.org/w/index.php?title=File:Pierre-Simon_Laplace.jpg *License:* Public Domain *Contributors:* Ashill, Ecummenic, Elcobbola, Gene.arboit, Jimmy44, Leye, Olivier, 霧木諒二, 1 anonymous edits

Image:student t pdf.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Student_t_pdf.svg *License:* Creative Commons Attribution 3.0 *Contributors:* Skbkckas

Image:student t cdf.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Student_t_cdf.svg *License:* Creative Commons Attribution 3.0 *Contributors:* Skbkckas

Image:T distribution 1df enhanced.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:T_distribution_1df_enhanced.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:IkamusumeFan

Image:T distribution 2df enhanced.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:T_distribution_2df_enhanced.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:IkamusumeFan

Image:T distribution 3df enhanced.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:T_distribution_3df_enhanced.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:IkamusumeFan

Image:T distribution 5df enhanced.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:T_distribution_5df_enhanced.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:IkamusumeFan

Image:T distribution 10df enhanced.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:T_distribution_10df_enhanced.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:IkamusumeFan

Image:T distribution 30df enhanced.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:T_distribution_30df_enhanced.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:IkamusumeFan

Image:F distributionPDF.png *Source:* https://en.wikipedia.org/w/index.php?title=File:F_distributionPDF.png *License:* GNU Free Documentation License *Contributors:* en:User:Pdbailey

Image:F distributionCDF.png *Source:* https://en.wikipedia.org/w/index.php?title=File:F_distributionCDF.png *License:* GNU Free Documentation License *Contributors:* en:User:Pdbailey

Image:Correlation examples2.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Correlation_examples2.svg *License:* Creative Commons Zero *Contributors:* DenisBoigelot, original uploader was Imagecreator

Image:correlation range dependence.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Correlation_range_dependence.svg *License:* Creative Commons Attribution 3.0 *Contributors:* Skbkckas

Image:Anscombe's quartet 3.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Anscombe's_quartet_3.svg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Anscombe.svg; Schutz derivative work (label using subscripts): Avenue (talk)

file:linear regression.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Linear_regression.svg *License:* Public Domain *Contributors:* Sewaqu

Image:Linear regression.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Linear_regression.svg *License:* Public Domain *Contributors:* Sewaqu

Image:Path example.JPG *Source:* https://en.wikipedia.org/w/index.php?title=File:Path_example.JPG *License:* Public Domain *Contributors:* Yunger

File:MovingAverage.GIF *Source:* <https://en.wikipedia.org/w/index.php?title=File:MovingAverage.GIF> *License:* Public Domain *Contributors:* Victory

File:Moving Average Types comparison - Simple and Exponential.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Moving_Average_Types_comparison_-_Simple_and_Exponential.png *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* Alex Kofman

Image:Weighted moving average weights N=15.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Weighted_moving_average_weights_N=15.png *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* Jochen Burghardt, Joxemai, Kevin Ryde

Image:Exponential moving average weights N=15.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Exponential_moving_average_weights_N=15.png *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* Jochen Burghardt, Joxemai, Kevin Ryde

File:ANOVA no fit.jpg *Source:* https://en.wikipedia.org/w/index.php?title=File:ANOVA_no_fit.jpg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Vanderlindenma

File:ANOVA fair fit.jpg *Source:* https://en.wikipedia.org/w/index.php?title=File:ANOVA_fair_fit.jpg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Vanderlindenma

File:ANOVA very good fit.jpg *Source:* https://en.wikipedia.org/w/index.php?title=File:ANOVA_very_good_fit.jpg *License:* Creative Commons Attribution-Sharealike 3.0 *Contributors:* User:Vanderlindenma

File:GaussianScatterPCA.png *Source:* <https://en.wikipedia.org/w/index.php?title=File:GaussianScatterPCA.png> *License:* GNU Free Documentation License *Contributors:* —Ben FrantzDale (talk) (Transferred by ILCyborg)

File:PCA of Haplogroup J using 37 STRs.png *Source:* https://en.wikipedia.org/w/index.php?title=File:PCA_of_Haplogroup_J_using_37_STRs.png *License:* Public Domain *Contributors:* User:Jheald

File:Elmap breastcancer wiki.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Elmap_breastcancer_wiki.png *License:* Public Domain *Contributors:* self-made, Андрей Зиновьев=Andrei Zinovyev

Image:Clusters.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Clusters.svg> *License:* Public Domain *Contributors:* Clusters.PNG; derivative work: Snubcube (talk)

Image:Hierarchical clustering simple diagram.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Hierarchical_clustering_simple_diagram.svg *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* File:Hierarchical_clustering_diagram.png#file: Stathis Sideris on 10/02/2005 derivative work: Mbbrugman (talk)

File:Linear-svm-scatterplot.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Linear-svm-scatterplot.svg> *License:* Creative Commons Zero *Contributors:* User:Qwertusy

File:Internet map 1024.jpg *Source:* https://en.wikipedia.org/w/index.php?title=File:Internet_map_1024.jpg *License:* Creative Commons Attribution 2.5 *Contributors:* Barrett Lyon The Opte Project

Image:K Means Example Step 1.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:K_Means_Example_Step_1.svg *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* Weston.pace

Image:K Means Example Step 2.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:K_Means_Example_Step_2.svg *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* Weston.pace

Image:K Means Example Step 3.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:K_Means_Example_Step_3.svg *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* Weston.pace

Image:K Means Example Step 4.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:K_Means_Example_Step_4.svg *License:* Creative Commons Attribution-ShareAlike 3.0 Unported *Contributors:* Weston.pace

File:K-means convergence to a local minimum.png *Source:* https://en.wikipedia.org/w/index.php?title=File:K-means_convergence_to_a_local_minimum.png *License:* Creative Commons Attribution-ShareAlike 3.0 *Contributors:* User:Agor153

File:Iris Flowers Clustering kMeans.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Iris_Flowers_Clustering_kMeans.svg *License:* Public Domain *Contributors:* Chire

File:ClusterAnalysis Mouse.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:ClusterAnalysis_Mouse.svg *License:* Public Domain *Contributors:* Chire

File:Rosa Gold Glow 2 small noblue.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Rosa_Gold_Glow_2_small_noblue.png *License:* GNU Free Documentation License *Contributors:* Dcoetzee

File:Rosa Gold Glow 2 small noblue color space.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Rosa_Gold_Glow_2_small_noblue_color_space.png *License:* Public Domain *Contributors:* User:Dcoetzee

File:Matrix.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Matrix.svg> *License:* Creative Commons Attribution-Share Alike *Contributors:* Lakeworks

file:Nuvola apps kaboodle.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Nuvola_apps_kaboodle.svg *License:* unknown *Contributors:* Cathy Richards, Pierpao, Tkgd2007, Waldir, 1 anonymous edits

File:Matrix multiplication diagram 2.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Matrix_multiplication_diagram_2.svg *License:* GNU Free Documentation License *Contributors:* User:Bilou

File:Area parallelogram as determinant.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Area_parallelogram_as_determinant.svg *License:* Public Domain *Contributors:* Jitse Niesen

File:VerticalShear m=1.25.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:VerticalShear_m=1.25.svg *License:* Public Domain *Contributors:* RobHar

File:Flip map.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Flip_map.svg *License:* Creative Commons Attribution-ShareAlike 3.0 *Contributors:* Jakob.schobach

File:Squeeze r=1.5.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Squeeze_r=1.5.svg *License:* Public Domain *Contributors:* RobHar

File:Scaling by 1.5.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Scaling_by_1.5.svg *License:* Public Domain *Contributors:* RobHar

File:Rotation by pi over 6.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Rotation_by_pi_over_6.svg *License:* Public Domain *Contributors:* RobHar

File:Ellipse in coordinate system with semi-axes labelled.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Ellipse_in_coordinate_system_with_semi-axes_labelled.svg *License:* Creative Commons Attribution-ShareAlike 3.0 *Contributors:* Jakob.schobach

File:Hyperbola2 SVG.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Hyperbola2_SVG.svg *License:* Creative Commons Attribution-ShareAlike 3.0 *Contributors:* User:IkamusumeFan

File:Determinant example.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Determinant_example.svg *License:* Creative Commons Attribution-ShareAlike 3.0 *Contributors:* User:Krishnavedala

File:Jordan blocks.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Jordan_blocks.svg *License:* Creative Commons Attribution-ShareAlike 3.0 *Contributors:* Jakob.schobach

File:Labelled undirected graph.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Labelled_undirected_graph.svg *License:* Creative Commons Attribution-ShareAlike 3.0 *Contributors:* Jakob.schobach

File:Saddle Point SVG.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Saddle_Point_SVG.svg *License:* Creative Commons Attribution-ShareAlike 3.0 *Contributors:* User:IkamusumeFan

File:Markov chain SVG.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Markov_chain_SVG.svg *License:* Creative Commons Attribution-ShareAlike 3.0 *Contributors:* User:IkamusumeFan

File:Matrix multiplication row column correspondance.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Matrix_multiplication_row_column_correspondance.svg *License:* Public Domain *Contributors:* User:Maschen

File:Bound on matrix multiplication omega over time.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Bound_on_matrix_multiplication_omega_over_time.svg *License:* Creative Commons Zero *Contributors:* Self

File:Block matrix multiplication.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Block_matrix_multiplication.svg *License:* Creative Commons Zero *Contributors:* User:Dcoetzee

File:Matrix transpose.gif *Source:* https://en.wikipedia.org/w/index.php?title=File:Matrix_transpose.gif *License:* Public Domain *Contributors:* Kieff

Image:Area parallelogram as determinant.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Area_parallelogram_as_determinant.svg *License:* Public Domain *Contributors:* Jitse Niesen

Image:Determinant parallelepiped.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Determinant_parallelepiped.svg *License:* Creative Commons Attribution 3.0 *Contributors:* Claudio Rocchini

File:Mona Lisa eigenvector grid.png *Source:* https://en.wikipedia.org/w/index.php?title=File:Mona_Lisa_eigenvector_grid.png *License:* Creative Commons Zero *Contributors:* TreyGreer62

File:Eigenvalue equation.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Eigenvalue_equation.svg *License:* Creative Commons Attribution-Share Alike *Contributors:* Lyudmil Antonov Lantonov 16:35, 13 March 2008 (UTC)

File:Eigenvectors.gif *Source:* <https://en.wikipedia.org/w/index.php?title=File:Eigenvectors.gif> *License:* Public Domain *Contributors:* Kieff

File:Homothety in two dim.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Homothety_in_two_dim.svg *License:* Creative Commons Attribution-Share Alike *Contributors:* Lyudmil Antonov -Lantonov 16:36, 13 March 2008 (UTC)

File:Unequal scaling.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Unequal_scaling.svg *License:* Creative Commons Attribution-Share Alike *Contributors:* Lyudmil Antonov -Lantonov 16:37, 13 March 2008 (UTC)

File:Rotation.png *Source:* <https://en.wikipedia.org/w/index.php?title=File:Rotation.png> *License:* Public Domain *Contributors:* Underdark

File:Shear.svg *Source:* <https://en.wikipedia.org/w/index.php?title=File:Shear.svg> *License:* Creative Commons Attribution-Share Alike *Contributors:* Lyudmil Antonov -Lantonov 09:13, 17 March 2008 (UTC)

File:HAtomOrbitals.png *Source:* <https://en.wikipedia.org/w/index.php?title=File:HAtomOrbitals.png> *License:* GNU Free Documentation License *Contributors:* Admrboltz, Benjah-bmm27, Dbc334, Dbenbenn, Ejdzaj, Falcorian, Hongsy, Kborland, MichaelDiederich, Mion, Saperaud, 6 anonymous edits

File:beam mode 1.gif *Source:* https://en.wikipedia.org/w/index.php?title=File:Beam_mode_1.gif *License:* GNU Free Documentation License *Contributors:* Original uploader was Lzyvzl at en.wikipedia

File:Eigenfaces.png *Source:* <https://en.wikipedia.org/w/index.php?title=File:Eigenfaces.png> *License:* Attribution *Contributors:* Laurascudder, Liftarn, Man vyi, Ylebru

Image:Secretsharing-3-point.png *Source:* <https://en.wikipedia.org/w/index.php?title=File:Secretsharing-3-point.png> *License:* GNU Free Documentation License *Contributors:* stib

Image:Intersecting Lines.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Intersecting_Lines.svg *License:* Public Domain *Contributors:* Jim.belk

Image:IntersectingPlanes.png *Source:* <https://en.wikipedia.org/w/index.php?title=File:IntersectingPlanes.png> *License:* GNU Free Documentation License *Contributors:* Original uploader was Stib at en.wikipedia

Image:One Line.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:One_Line.svg *License:* Public Domain *Contributors:* Jim.belk

Image:Two Lines.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Two_Lines.svg *License:* Public Domain *Contributors:* Jim.belk, Krishnavedala, Sarang, 2 anonymous edits

Image:Three Lines.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Three_Lines.svg *License:* Public Domain *Contributors:* Jim.belk

Image:Three Intersecting Lines.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Three_Intersecting_Lines.svg *License:* Public Domain *Contributors:* Jim.belk

Image:Parallel Lines.svg *Source:* https://en.wikipedia.org/w/index.php?title=File:Parallel_Lines.svg *License:* Public Domain *Contributors:* Jim.belk

License

Creative Commons Attribution-Share Alike 3.0
[//creativecommons.org/licenses/by-sa/3.0/](https://creativecommons.org/licenses/by-sa/3.0/)
