

Normal distribution

In probability theory, the **normal** (or **Gaussian** or **Gauss** or **Laplace–Gauss**) **distribution** is a very common continuous probability distribution. Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known.^{[1][2]} A random variable with a Gaussian distribution is said to be **normally distributed** and is called a **normal deviate**.

The normal distribution is useful because of the central limit theorem. In its most general form, under some conditions (which include finite variance), it states that averages of samples of observations of random variables independently drawn from the same distribution converge in distribution to the normal, that is, they become normally distributed when the number of observations is sufficiently large. Physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have distributions that are nearly normal.^[3] 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 normal distribution is sometimes informally called the **bell curve**. However, many other distributions are bell-shaped (such as the Cauchy, Student's *t*-, and logistic distributions).

The probability density of the normal distribution is

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

where

- μ is the mean or expectation of the distribution (and also its median and mode)
- σ is the standard deviation
- σ^2 is the variance

Contents

Definition

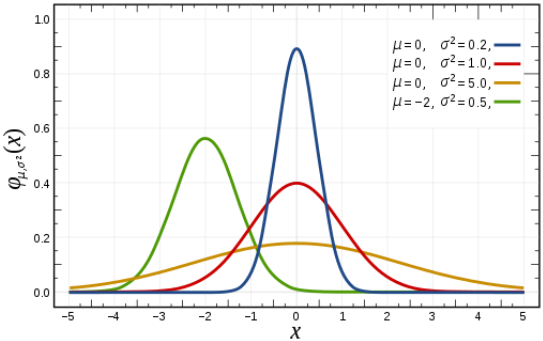
- Standard normal distribution
- General normal distribution
- Notation
- Alternative parameterizations

Properties

- Symmetries and derivatives
- Moments
- Fourier transform and characteristic function
- Moment and cumulant generating functions

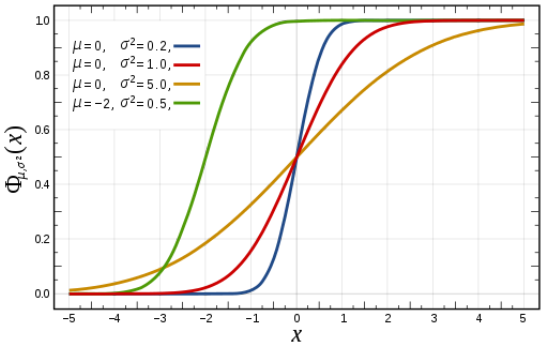
Normal distribution

Probability density function



The red curve is the *standard normal distribution*

Cumulative distribution function



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \right]$

- Cumulative distribution function**
 - Standard deviation and coverage
 - Quantile function
- Zero-variance limit**
- Central limit theorem**
- Maximum entropy**
- Operations on normal deviates**
 - Infinite divisibility and Cramér's theorem
 - Bernstein's theorem
- Other properties**
- Related distributions**
 - Operations on a single random variable
 - Combination of two independent random variables
 - Combination of two or more independent random variables
 - Operations on the density function
 - Extensions
- Normality tests**
- Estimation of parameters**
 - Sample mean
 - Sample variance
 - Confidence intervals
- Bayesian analysis of the normal distribution**
 - Sum of two quadratics
 - Scalar form
 - Vector form
 - Sum of differences from the mean
 - With known variance
 - With known mean
 - With unknown mean and unknown variance
- Occurrence and applications**
 - Exact normality
 - Approximate normality
 - Assumed normality
 - Produced normality
- Generating values from normal distribution**
- Numerical approximations for the normal CDF**
- History**
 - Development
 - Naming
- See also**
- Notes**
- References**
 - Citations
 - Sources

Quantile	 μ<!-- μ --> + σ<!-- σ --> 2 erf −<!-- − --> 1 (2 p −<!-- − --> 1) {\displaystyle \mu +\sigma {\sqrt {2}}\mathrm {erf} ^{-1}(2p-1)}
Mean	 μ<!-- μ --> {\displaystyle \mu }
Median	 μ<!-- μ --> {\displaystyle \mu }
Mode	 μ<!-- μ --> {\displaystyle \mu }
Variance	 σ<!-- σ --> 2 {\displaystyle \sigma ^{2}}
Skewness	0
Ex. kurtosis	0
Entropy	 1 2 log ⁡<!-- ⁡ --> (2 π<!-- π --> e σ<!-- σ --> 2) {\displaystyle {\frac {1}{2}}\log(2\pi e\sigma ^{2})}
MGF	 exp ⁡<!-- ⁡ --> (μ<!-- μ --> t + σ<!-- σ --> 2 t 2 / 2) {\displaystyle \exp(\mu t+\sigma ^{2}t^{2}/2)}
CF	 exp ⁡<!-- ⁡ --> (i μ<!-- μ --> t −<!-- − --> σ<!-- σ --> 2 t 2 / 2) {\displaystyle \exp(i\mu t-\sigma ^{2}t^{2}/2)}
Fisher information	<div> I (μ<!-- μ --> , σ<!-- σ -->) = (1 σ<!-- σ --> 2 0 0 2 / σ<!-- σ --> 2) {\displaystyle {\mathcal {I}}(\mu ,\sigma)={\begin{pmatrix}1/\sigma ^{2}&0\\0&2/\sigma ^{2}\end{pmatrix}} </div> <div> I (μ<!-- μ --> , σ<!-- σ --> 2) = (1 σ<!-- σ --> 2 0 0 1 / (2 σ<!-- σ --> 4)) {\displaystyle {\mathcal {I}}(\mu ,\sigma ^{2})={\begin{pmatrix}1/\sigma ^{2}&0\\0&1/(2\sigma ^{4})\end{pmatrix}} </div>
Kullback-Leibler divergence	 D KL (N<!-- N --> 0 ∥<!-- ∥ --> N<!-- N --> 1) = 1 2 { (σ<!-- σ --> 0 / σ<!-- σ --> 1) 2 + (μ<!-- μ --> 1 −<!-- − --> μ<!-- μ --> 0) 2 σ<!-- σ --> 1 2 −<!-- − --> 1 + 2 ln ⁡<!-- ⁡ --> σ<!-- σ --> 1 σ<!-- σ --> 0 } {\displaystyle D_{\mathrm {KL} }({\mathcal {N}}_{0}\ {\mathcal {N}}_{1})={\frac {1}{2}}\{(\sigma _{0}/\sigma _{1})^{2}+{\frac {(\mu _{1}-\mu _{0})^{2}}{\sigma _{1}^{2}}}-1+2\ln {\frac {\sigma _{1}}{\sigma _{0}}}\}

External links

Definition

Standard normal distribution

The simplest case of a normal distribution is known as the *standard normal distribution*. This is a special case when $\mu = 0$ and $\sigma = 1$, and it is described by this probability density function:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

The factor $1/\sqrt{2\pi}$ in this expression ensures that the total area under the curve $\varphi(x)$ is equal to one.^[note 1] The factor $1/2$ in the exponent ensures that the distribution has unit variance (i.e. the variance is equal to one), 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 $x = +1$ and $x = -1$.

Authors differ on which normal distribution should be called the "standard" one. Gauss defined the standard normal as having variance $\sigma^2 = 1/2$, that is

$$\varphi(x) = \frac{e^{-x^2}}{\sqrt{\pi}}.$$

Stigler^[4] goes even further, defining the standard normal with variance $\sigma^2 = 1/(2\pi)$:

$$\varphi(x) = e^{-\pi x^2}.$$

General normal distribution

Every 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):

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}\right).$$

The probability density must be scaled by $1/\sigma$ so that the integral is still 1.

If Z is a standard normal deviate, then $X = \sigma Z + \mu$ will have a normal distribution with expected value μ and standard deviation σ . Conversely, if X is a normal deviate with parameters μ and σ^2 , then $Z = (X - \mu)/\sigma$ will have a standard normal distribution. This variate is called the standardized form of X .

Every normal distribution is the exponential of a quadratic function:

$$f(x) = e^{ax^2 + bx + c}$$

where $a < 0$ and $c = b^2/(4a) + \ln(-a/\pi)/2$. In this form, the mean value is $\mu = -b/(2a)$, and the variance is $\sigma^2 = -1/(2a)$. For the standard normal distribution, $a = -1/2$, $b = 0$, and $c = -\ln(2\pi)/2$.

Notation

The probability density of the standard Gaussian distribution (standard normal distribution) (with zero mean and unit variance) is often denoted with the Greek letter ϕ (phi).^[5] The alternative form of the Greek letter phi, φ , is also used quite often.

The normal distribution is often referred to as $N(\mu, \sigma^2)$ or $\mathcal{N}(\mu, \sigma^2)$.^[6] Thus when a random variable X is distributed normally with mean μ and variance σ^2 , one may write

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

Alternative parameterizations

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, $1/\sigma^2$.^[7] The formula for the distribution then becomes

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

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.

Also the reciprocal of the standard deviation $\tau' = 1/\sigma$ might be defined as the *precision* and the expression of the normal distribution becomes

$$f(x) = \frac{\tau'}{\sqrt{2\pi}} e^{-(\tau')^2(x-\mu)^2/2}.$$

According to Stigler, this formulation is advantageous because of a much simpler and easier-to-remember formula, and simple approximate formulas for the quantiles of the distribution.

Normal distributions form 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 $\eta_2 = \mu^2 + \sigma^2$.

Properties

The normal distribution is the only absolutely continuous distribution 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 specified mean and variance.^{[8][9]} Geary has shown, assuming that the mean and variance are finite, that the normal distribution is the only distribution where the mean and variance calculated from a set of independent draws are independent of each other.^{[10][11]}

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 (e.g., a spread of three standard deviations covers all but 0.27% of the total distribution). 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, a more heavy-tailed distribution should be assumed and the appropriate robust statistical inference methods applied.

The Gaussian distribution belongs to the family of stable distributions which are the attractors of sums of independent, identically distributed distributions whether or not the mean or variance is finite. Except for the Gaussian which is a limiting case, all stable distributions have heavy tails and infinite variance. It is one of the few distributions that are stable and that have probability density functions that can be expressed analytically, the others being the Cauchy distribution and the Lévy distribution.

Symmetries and derivatives

The normal distribution with density $f(\mathbf{x})$ (mean $\boldsymbol{\mu}$ and standard deviation $\sigma > 0$) has the following properties:

- It is symmetric around the point $\mathbf{x} = \boldsymbol{\mu}$, which is at the same time the mode, the median and the mean of the distribution.^[12]
- It is unimodal: its first derivative is positive for $\mathbf{x} < \boldsymbol{\mu}$, negative for $\mathbf{x} > \boldsymbol{\mu}$, and zero only at $\mathbf{x} = \boldsymbol{\mu}$.
- The area under the curve and over the \mathbf{x} -axis is unity (i.e. equal to one).
- Its density 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 $\mathbf{x} = \boldsymbol{\mu} - \sigma$ and $\mathbf{x} = \boldsymbol{\mu} + \sigma$.^[12]
- Its density is log-concave.^[12]
- Its density is infinitely differentiable, indeed supersmooth of order 2.^[13]

Furthermore, the density φ of the standard normal distribution (i.e. $\boldsymbol{\mu} = \mathbf{0}$ and $\sigma = 1$) also has the following properties:

- Its first derivative is $\varphi'(\mathbf{x}) = -\mathbf{x}\varphi(\mathbf{x})$.
- Its second derivative is $\varphi''(\mathbf{x}) = (\mathbf{x}^2 - 1)\varphi(\mathbf{x})$
- More generally, its n th derivative is $\varphi^{(n)}(\mathbf{x}) = (-1)^n \mathbf{He}_n(\mathbf{x})\varphi(\mathbf{x})$, where $\mathbf{He}_n(\mathbf{x})$ is the n th (probabilist) Hermite polynomial.^[14]
- The probability that a normally distributed variable \mathbf{X} with known $\boldsymbol{\mu}$ and σ is in a particular set, can be calculated by using the fact that the fraction $\mathbf{Z} = (\mathbf{X} - \boldsymbol{\mu})/\sigma$ has a standard normal distribution.

Moments

The plain and absolute moments of a variable \mathbf{X} are the expected values of \mathbf{X}^p and $|\mathbf{X}|^p$, respectively. If the expected value $\boldsymbol{\mu}$ of \mathbf{X} is zero, these parameters are called *central moments*. Usually we are interested only in moments with integer order p .

If \mathbf{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:^[15]

$$\mathbf{E}[\mathbf{X}^p] = \begin{cases} 0 & \text{if } p \text{ is odd,} \\ \sigma^p (p-1)!! & \text{if } p \text{ is even.} \end{cases}$$

Here $n!!$ denotes the double factorial, that is, the product of all numbers from n to 1 that have the same parity as n .

The central absolute moments coincide with plain moments for all even orders, but are nonzero for odd orders. For any non-negative integer p ,

$$\mathbf{E}[|\mathbf{X}|^p] = \sigma^p (p-1)!! \cdot \begin{cases} \sqrt{\frac{2}{\pi}} & \text{if } p \text{ is odd} \\ 1 & \text{if } p \text{ is even} \end{cases} = \sigma^p \cdot \frac{2^{p/2} \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$$

The last formula is valid also for any non-integer $p > -1$. When the mean $\boldsymbol{\mu} \neq \mathbf{0}$, the plain and absolute moments can be expressed in terms of confluent hypergeometric functions ${}_1F_1$ and U .

$$\begin{aligned} \mathbf{E}[\mathbf{X}^p] &= \sigma^p \cdot (-i\sqrt{2})^p U\left(-\frac{p}{2}, \frac{1}{2}, -\frac{1}{2}\left(\frac{\boldsymbol{\mu}}{\sigma}\right)^2\right), \\ \mathbf{E}[|\mathbf{X}|^p] &= \sigma^p \cdot 2^{p/2} \frac{\Gamma\left(\frac{1+p}{2}\right)}{\sqrt{\pi}} {}_1F_1\left(-\frac{p}{2}, \frac{1}{2}, -\frac{1}{2}\left(\frac{\boldsymbol{\mu}}{\sigma}\right)^2\right). \end{aligned}$$

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	$\mu^8 + 28\mu^6\sigma^2 + 210\mu^4\sigma^4 + 420\mu^2\sigma^6 + 105\sigma^8$	$105\sigma^8$

The expectation of \boldsymbol{X} conditioned on the event that \boldsymbol{X} lies in an interval $[\boldsymbol{a}, \boldsymbol{b}]$ is given by

$$\mathbf{E}[\boldsymbol{X} \mid \boldsymbol{a} < \boldsymbol{X} < \boldsymbol{b}] = \mu - \sigma^2 \frac{f(\boldsymbol{b}) - f(\boldsymbol{a})}{F(\boldsymbol{b}) - F(\boldsymbol{a})}$$

where \boldsymbol{f} and \boldsymbol{F} respectively are the density and the cumulative distribution function of \boldsymbol{X} . For $\boldsymbol{b} = \infty$ this is known as the inverse Mills ratio. Note that above, density \boldsymbol{f} of \boldsymbol{X} is used instead of standard normal density as in inverse Mills ratio, so here we have σ^2 instead of σ .

Fourier transform and characteristic function

The Fourier transform of a normal density \boldsymbol{f} with mean μ and standard deviation σ is^[16]

$$\hat{f}(t) = \int_{-\infty}^{\infty} f(x)e^{-itx} \, dx = e^{-i\mu t} e^{-\frac{1}{2}(\sigma t)^2}$$

where \boldsymbol{i} is the imaginary unit. If the mean $\mu = \mathbf{0}$, the first factor is 1, and the Fourier transform is, apart from a constant factor, a normal density on the frequency domain, with mean 0 and standard deviation $\mathbf{1}/\sigma$. In particular, the standard normal distribution φ is an eigenfunction of the Fourier transform.

In probability theory, the Fourier transform of the probability distribution of a real-valued random variable \boldsymbol{X} is closely connected to the characteristic function $\varphi_{\boldsymbol{X}}(\boldsymbol{t})$ of that variable, which is defined as the expected value of $e^{i\boldsymbol{t}\boldsymbol{X}}$, as a function of the real variable \boldsymbol{t} (the frequency parameter of the Fourier transform). This definition can be analytically extended to a complex-value variable \boldsymbol{t} .^[17] The relation between both is:

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \hat{f}(-\boldsymbol{t})$$

Moment and cumulant generating functions

The moment generating function of a real random variable \boldsymbol{X} is the expected value of $e^{\boldsymbol{t}\boldsymbol{X}}$, as a function of the real parameter \boldsymbol{t} . For a normal distribution with density \boldsymbol{f} , mean μ and deviation σ , the moment generating function exists and is equal to

$$\boldsymbol{M}(\boldsymbol{t}) = \mathbf{E}[e^{\boldsymbol{t}\boldsymbol{X}}] = \hat{f}(i\boldsymbol{t}) = e^{\mu \boldsymbol{t}} e^{\frac{1}{2}\sigma^2 \boldsymbol{t}^2}$$

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 function

The cumulative distribution function (CDF) of the standard normal distribution, usually denoted with the capital Greek letter Φ (phi), is the integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

The related error function **erf**(x) gives the probability of a random variable with normal distribution of mean 0 and variance 1/2 falling in the range $[-x, x]$; that is

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

These integrals cannot be expressed in terms of elementary functions, and are often said to be special functions. However, many numerical approximations are known; see below.

The two functions 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 with density f , 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.^{[18][19]} It gives the probability that the value of a standard normal random variable X will exceed x : $P(X > x)$. Other definitions of the Q -function, all of which are simple transformations of Φ , are also used occasionally.^[20]

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) is

$$\int \Phi(x) dx = x\Phi(x) + \varphi(x) + C.$$

The CDF of the standard normal distribution can be expanded by Integration by parts into a series:

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \left[x + \frac{x^3}{3} + \frac{x^5}{3 \cdot 5} + \cdots + \frac{x^{2n+1}}{(2n+1)!!} + \cdots \right]$$

where !! denotes the double factorial.

An asymptotic expansion of the CDF for large x can also be derived using integration by parts; see Error function#Asymptotic expansion.^[21]

Standard deviation and coverage

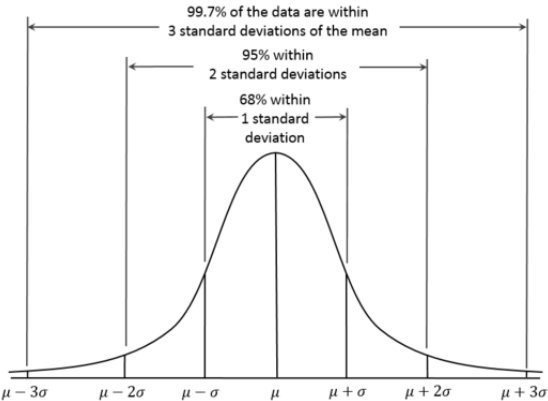
About 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations. This fact is known as the 68-95-99.7 (empirical) rule, or the *3-sigma rule*.

More precisely, the probability that a normal deviate lies in the range between $\mu - n\sigma$ 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 significant figures, the values for $n = 1, 2, \dots, 6$ are:^[22]

<i>n</i>	<i>p</i> = <i>F</i> ($\mu + n\sigma$) − <i>F</i> ($\mu − n\sigma$)	i.e. 1 − <i>p</i>	or 1 in <i>p</i>	OEIS
1	0.682 689 492 137	0.317 310 507 863	3.151 487 187 53	OEIS: A178647
2	0.954 499 736 104	0.045 500 263 896	21.977 894 5080	OEIS: A110894
3	0.997 300 203 937	0.002 699 796 063	370.398 347 345	OEIS: A270712
4	0.999 936 657 516	0.000 063 342 484	15 787.192 7673	
5	0.999 999 426 697	0.000 000 573 303	1 744 277.893 62	
6	0.999 999 998 027	0.000 000 001 973	506 797 345.897	



For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%.

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 \boldsymbol{X} will exceed $\mu + z_p\sigma$ with probability $1 - p$, and will lie outside the interval $\mu \pm z_p\sigma$ 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 quantile z_p such that \boldsymbol{X} will lie in the range $\mu \pm z_p\sigma$ 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:^{[23][24]} NOTE: the following table shows $\sqrt{2} \operatorname{erf}^{-1}(p) = \Phi^{-1}\left(\frac{p+1}{2}\right)$, not $\Phi^{-1}(p)$ as defined above.

<i>p</i>	<i>z_p</i>	<i>p</i>	<i>z_p</i>
0.80	1.281 551 565 545	0.999	3.290 526 731 492
0.90	1.644 853 626 951	0.9999	3.890 591 886 413
0.95	1.959 963 984 540	0.99999	4.417 173 413 469
0.98	2.326 347 874 041	0.999999	4.891 638 475 699
0.99	2.575 829 303 549	0.9999999	5.326 723 886 384
0.995	2.807 033 768 344	0.99999999	5.730 728 868 236
0.998	3.090 232 306 168	0.999999999	6.109 410 204 869

For small *p*, the quantile function has the useful asymptotic expansion $\Phi^{-1}(p) = -\sqrt{\ln \frac{1}{p^2} - \ln \ln \frac{1}{p^2} - \ln(2\pi)} + o(1)$.

Zero-variance limit

In the limit when *σ* tends to zero, the probability density *f*(*x*) eventually tends to zero at any *x* ≠ *μ*, but grows without limit if *x* = *μ*, while its integral remains equal to 1. Therefore, the normal distribution cannot be defined as an ordinary function when *σ* = 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*) = *δ*(*x* − *μ*). Its CDF is then the Heaviside step function translated by the mean *μ*, namely

$$F(x) = \begin{cases} 0 & \text{if } x < \mu \\ 1 & \text{if } x \geq \mu \end{cases}$$

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, . . . , *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}

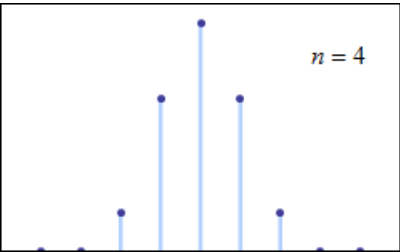
$$Z = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n X_i \right)$$

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:



As the number of discrete events increases, the function begins to resemble a normal distribution

- 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 0 or 1.
- The Poisson distribution with parameter λ is approximately normal with mean λ and variance λ , for large values of λ .^[25]
- The chi-squared distribution $\chi^2(k)$ is approximately normal with mean k and variance $2k$, for large k .
- The Student's t-distribution $t(\nu)$ is approximately normal with mean 0 and variance 1 when ν 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.

Maximum entropy

Of all probability distributions over the reals with a specified mean μ and variance σ^2 , the normal distribution $N(\mu, \sigma^2)$ is the one with maximum entropy.^[26] If \mathbf{X} is a continuous random variable with probability density $f(\mathbf{x})$, then the entropy of \mathbf{X} is defined as^{[27][28][29]}

$$H(\mathbf{X}) = - \int_{-\infty}^{\infty} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} = \frac{1}{2} (1 + \log(2\sigma^2\pi))$$

where $f(\mathbf{x}) \log f(\mathbf{x})$ is understood to be zero whenever $f(\mathbf{x}) = 0$. This functional can be maximized, subject to the constraints that the distribution is properly normalized and has a specified variance, by using variational calculus. A function with two Lagrange multipliers is defined:

$$L = \int_{-\infty}^{\infty} f(\mathbf{x}) \ln(f(\mathbf{x})) d\mathbf{x} - \lambda_0 \left(1 - \int_{-\infty}^{\infty} f(\mathbf{x}) d\mathbf{x} \right) - \lambda \left(\sigma^2 - \int_{-\infty}^{\infty} f(\mathbf{x})(\mathbf{x} - \mu)^2 d\mathbf{x} \right)$$

where $f(\mathbf{x})$ is, for now, regarded as some density function with mean μ and standard deviation σ .

At maximum entropy, a small variation $\delta f(\mathbf{x})$ about $f(\mathbf{x})$ will produce a variation δL about L which is equal to 0:

$$0 = \delta L = \int_{-\infty}^{\infty} \delta f(\mathbf{x}) (\ln(f(\mathbf{x})) + 1 + \lambda_0 + \lambda(\mathbf{x} - \mu)^2) d\mathbf{x}$$

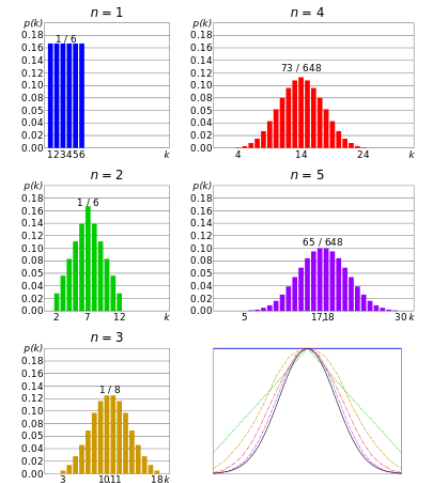
Since this must hold for any small $\delta f(\mathbf{x})$, the term in brackets must be zero, and solving for $f(\mathbf{x})$ yields:

$$f(\mathbf{x}) = e^{-\lambda_0 - 1 - \lambda(\mathbf{x} - \mu)^2}$$

Using the constraint equations to solve for λ_0 and λ yields the density of the normal distribution:

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

Operations on normal deviates



Comparison of probability density functions, $p(k)$ for the sum of n fair 6-sided dice to show their convergence to a normal distribution with increasing n , in accordance to the central limit theorem. In the bottom-right graph, smoothed profiles of the previous graphs are rescaled, superimposed and compared with a normal distribution (black curve).

The family of normal distributions is closed under linear transformations: if X is normally distributed with mean μ and standard 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 standard 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.^[30]

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.^[31]

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.^[32]

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 closely.^[33]

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.^{[34][35]}

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 denotes the variance of X_k .^[34]

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 is a normal random variable.^[33] 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.^{[36][37][proof]} For non-normal random variables uncorrelatedness does not imply independence.
3. The Kullback–Leibler divergence of one normal distribution $X_1 \sim N(\mu_1, \sigma_1^2)$ from another $X_2 \sim N(\mu_2, \sigma_2^2)$ is given by:^[38]

$$D_{\text{KL}}(X_1 \parallel X_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} + \frac{1}{2} \left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \ln \frac{\sigma_1^2}{\sigma_2^2} \right).$$

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. The conjugate prior of the mean of a normal distribution is another normal distribution.^[39] 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

$$\mu \mid x_1, \dots, x_n \sim \mathcal{N}\left(\frac{\frac{\sigma^2}{n}\mu_0 + \sigma_0^2\bar{x}}{\frac{\sigma^2}{n} + \sigma_0^2}, \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\right)$$

6. The family of normal distributions not only forms an exponential family (EF), but in fact forms a natural exponential family (NEF) with quadratic variance function (NEF-QVF). Many properties of normal distributions generalize to properties of NEF-QVF distributions, NEF distributions, or EF distributions generally. NEF-QVF distributions comprises 6 families, including Poisson, Gamma, binomial, and negative binomial distributions, while many of the common families studied in probability and statistics are NEF or EF.

7. In information geometry, the family of normal distributions forms a statistical manifold with constant curvature -1 . The same family is flat with respect to the (± 1) -connections $\nabla^{(e)}$ and $\nabla^{(m)}$.^[40]

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 absolute value of normalized residuals, $|X - \mu|/\sigma$, has chi distribution with one degree of freedom: $|X - \mu|/\sigma \sim \chi_1(|X - \mu|/\sigma)$.
- The square of X/σ has the noncentral chi-squared distribution with one degree of freedom: $X^2/\sigma^2 \sim \chi_1^2(\mu^2/\sigma^2)$. 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^[41] 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: $X_1 / X_2 \sim \text{Cauchy}(0, 1)$.
- 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

$$X_1^2 + \cdots + X_n^2 \sim \chi_n^2.$$

- 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,^[42] which can be demonstrated using Basu's theorem or Cochran's theorem.^[43] The ratio of these two quantities will have the Student's t-distribution with $n - 1$ degrees of freedom:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\frac{1}{n}(X_1 + \cdots + X_n) - \mu}{\sqrt{\frac{1}{n(n-1)} \left[(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2 \right]}} \sim t_{n-1}.$$

- 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:^[44]

$$F = \frac{(X_1^2 + X_2^2 + \cdots + X_n^2)/n}{(Y_1^2 + Y_2^2 + \cdots + Y_m^2)/m} \sim F_{n,m}.$$

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.

A random variable X has a two-piece normal distribution if it has a distribution

$$\begin{aligned} f_X(x) &= N(\mu, \sigma_1^2) \text{ if } x \leq \mu \\ f_X(x) &= N(\mu, \sigma_2^2) \text{ if } x \geq \mu \end{aligned}$$

where μ is the mean and σ_1 and σ_2 are the standard deviations of the distribution to the left and right of the mean respectively.

The mean, variance and third central moment of this distribution have been determined^[45]

$$\mathbf{E}(X) = \mu + \sqrt{\frac{2}{\pi}}(\sigma_2 - \sigma_1)$$

$$\mathbf{V}(X) = \left(1 - \frac{2}{\pi}\right)(\sigma_2 - \sigma_1)^2 + \sigma_1\sigma_2$$

$$\mathbf{T}(X) = \sqrt{\frac{2}{\pi}}(\sigma_2 - \sigma_1) \left[\left(\frac{4}{\pi} - 1\right)(\sigma_2 - \sigma_1)^2 + \sigma_1\sigma_2 \right]$$

where $\mathbf{E}(X)$, $\mathbf{V}(X)$ and $\mathbf{T}(X)$ are the mean, variance, and third central moment respectively.

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-parameter family of probability distributions that extend the normal law to include different skewness and kurtosis values.
- The generalized normal distribution, also known as the exponential power distribution, allows for distribution tails with thicker or thinner asymptotic behaviors.

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*:

$$\hat{\mu} = \bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Sample mean

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.^[46] 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:

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Sample variance

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),^[46] 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 .

Confidence intervals

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 converse 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 ;^[47] similarly, inverting the χ^2 distribution of the statistic s^2 will give us the confidence interval for σ^2 .^[48]

$$\begin{aligned} \mu &\in \left[\hat{\mu} - t_{n-1, 1-\alpha/2} \frac{1}{\sqrt{n}} s, \hat{\mu} + t_{n-1, 1-\alpha/2} \frac{1}{\sqrt{n}} s \right] \approx \left[\hat{\mu} - |z_{\alpha/2}| \frac{1}{\sqrt{n}} s, \hat{\mu} + |z_{\alpha/2}| \frac{1}{\sqrt{n}} s \right], \\ \sigma^2 &\in \left[\frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2} \right] \approx \left[s^2 - |z_{\alpha/2}| \frac{\sqrt{2}}{\sqrt{n}} s^2, s^2 + |z_{\alpha/2}| \frac{\sqrt{2}}{\sqrt{n}} s^2 \right], \end{aligned}$$

where $t_{k,p}$ and $\chi_{k,p}^2$ are the p th quantiles of the t - and χ^2 -distributions respectively. These confidence intervals are of the *confidence level* $1 - \alpha$, meaning that the true values μ and σ^2 fall outside of these intervals with probability (or significance level) 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.

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

$$\begin{aligned} & (\mathbf{y} - \mathbf{x})' \mathbf{A} (\mathbf{y} - \mathbf{x}) + (\mathbf{x} - \mathbf{z})' \mathbf{B} (\mathbf{x} - \mathbf{z}) \\ &= (\mathbf{x} - \mathbf{c})' (\mathbf{A} + \mathbf{B}) (\mathbf{x} - \mathbf{c}) + (\mathbf{y} - \mathbf{z})' (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} (\mathbf{y} - \mathbf{z}) \end{aligned}$$

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:

$$\mathbf{x}' \mathbf{A} \mathbf{x} = \sum_{i,j} a_{ij} x_i x_j$$

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 $x_i x_j = x_j x_i$, 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 $\mathbf{x}' \mathbf{A} \mathbf{y} = \mathbf{y}' \mathbf{A} \mathbf{x}$.

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

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 $x \sim \mathcal{N}(\mu, 1/\tau)$ and $\mu \sim \mathcal{N}(\mu_0, 1/\tau_0)$, we proceed as follows.

First, the likelihood function is (using the formula above for the sum of differences from the mean):

$$\begin{aligned} p(\mathbf{X} \mid \mu, \tau) &= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x_i - \mu)^2\right) \\ &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\tau \sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{1}{2}\tau \left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)\right]. \end{aligned}$$

Then, we proceed as follows:

$$\begin{aligned} p(\mu \mid \mathbf{X}) &\propto p(\mathbf{X} \mid \mu)p(\mu) \\ &= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{1}{2}\tau \left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)\right] \sqrt{\frac{\tau_0}{2\pi}} \exp\left(-\frac{1}{2}\tau_0(\mu - \mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\tau \left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right) + \tau_0(\mu - \mu_0)^2\right)\right) \\ &\propto \exp\left(-\frac{1}{2}(n\tau(\bar{x} - \mu)^2 + \tau_0(\mu - \mu_0)^2)\right) \\ &= \exp\left(-\frac{1}{2}(n\tau + \tau_0)\left(\mu - \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}\right)^2 + \frac{n\tau\tau_0}{n\tau + \tau_0}(\bar{x} - \mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2}(n\tau + \tau_0)\left(\mu - \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}\right)^2\right) \end{aligned}$$

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.

$$p(\mu \mid \mathbf{X}) \sim \mathcal{N}\left(\frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}, \frac{1}{n\tau + \tau_0}\right)$$

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 $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$ 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:

$$p(\sigma^2 \mid \nu_0, \sigma_0^2) = \frac{(\sigma_0^2 \frac{\nu_0}{2})^{\nu_0/2}}{\Gamma(\frac{\nu_0}{2})} \frac{\exp\left[\frac{-\nu_0\sigma_0^2}{2\sigma^2}\right]}{(\sigma^2)^{1+\frac{\nu_0}{2}}} \propto \frac{\exp\left[\frac{-\nu_0\sigma_0^2}{2\sigma^2}\right]}{(\sigma^2)^{1+\frac{\nu_0}{2}}}$$

The likelihood function from above, written in terms of the variance, is:

$$\begin{aligned}
 p(\mathbf{X} \mid \mu, \sigma^2) &= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] \\
 &= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left[-\frac{S}{2\sigma^2} \right]
 \end{aligned}$$

where

$$S = \sum_{i=1}^n (x_i - \mu)^2.$$

Then:

$$\begin{aligned}
 p(\sigma^2 \mid \mathbf{X}) &\propto p(\mathbf{X} \mid \sigma^2) p(\sigma^2) \\
 &= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left[-\frac{S}{2\sigma^2} \right] \frac{(\sigma_0^2 \frac{\nu_0}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})} \frac{\exp \left[\frac{-\nu_0 \sigma_0^2}{2\sigma^2} \right]}{(\sigma^2)^{1+\frac{\nu_0}{2}}} \\
 &\propto \left(\frac{1}{\sigma^2} \right)^{n/2} \frac{1}{(\sigma^2)^{1+\frac{\nu_0}{2}}} \exp \left[-\frac{S}{2\sigma^2} + \frac{-\nu_0 \sigma_0^2}{2\sigma^2} \right] \\
 &= \frac{1}{(\sigma^2)^{1+\frac{\nu_0+n}{2}}} \exp \left[-\frac{\nu_0 \sigma_0^2 + S}{2\sigma^2} \right]
 \end{aligned}$$

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:

$$p(\mu \mid \sigma^2; \mu_0, n_0) \sim \mathcal{N}(\mu_0, \sigma^2/n_0)$$

$$p(\sigma^2; \nu_0, \sigma_0^2) \sim I\chi^2(\nu_0, \sigma_0^2) = IG(\nu_0/2, \nu_0\sigma_0^2/2)$$

The update equations can be derived, and look as follows:

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

$$\mu'_0 = \frac{n_0\mu_0 + n\bar{x}}{n_0 + n}$$

$$n'_0 = n_0 + n$$

$$\nu'_0 = \nu_0 + n$$

$$\nu'_0\sigma_0'^2 = \nu_0\sigma_0^2 + \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n_0n}{n_0 + n}(\mu_0 - \bar{x})^2$$

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]

The prior distributions are

$$\begin{aligned} p(\mu \mid \sigma^2; \mu_0, n_0) &\sim \mathcal{N}(\mu_0, \sigma^2/n_0) = \frac{1}{\sqrt{2\pi \frac{\sigma^2}{n_0}}} \exp\left(-\frac{n_0}{2\sigma^2}(\mu - \mu_0)^2\right) \\ &\propto (\sigma^2)^{-1/2} \exp\left(-\frac{n_0}{2\sigma^2}(\mu - \mu_0)^2\right) \\ p(\sigma^2; \nu_0, \sigma_0^2) &\sim I\chi^2(\nu_0, \sigma_0^2) = IG(\nu_0/2, \nu_0 \sigma_0^2/2) \\ &= \frac{(\sigma_0^2 \nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \frac{\exp\left[\frac{-\nu_0 \sigma_0^2}{2\sigma^2}\right]}{(\sigma^2)^{1+\nu_0/2}} \\ &\propto (\sigma^2)^{-(1+\nu_0/2)} \exp\left[\frac{-\nu_0 \sigma_0^2}{2\sigma^2}\right]. \end{aligned}$$

Therefore, the joint prior is

$$\begin{aligned} p(\mu, \sigma^2; \mu_0, n_0, \nu_0, \sigma_0^2) &= p(\mu \mid \sigma^2; \mu_0, n_0) p(\sigma^2; \nu_0, \sigma_0^2) \\ &\propto (\sigma^2)^{-(\nu_0+3)/2} \exp\left[-\frac{1}{2\sigma^2}(\nu_0 \sigma_0^2 + n_0(\mu - \mu_0)^2)\right]. \end{aligned}$$

The likelihood function from the section above with known variance is:

$$p(\mathbf{X} \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \mu)^2\right)\right]$$

Writing it in terms of variance rather than precision, we get:

$$\begin{aligned} p(\mathbf{X} \mid \mu, \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)\right] \\ &\propto \sigma^{2-n/2} \exp\left[-\frac{1}{2\sigma^2} (S + n(\bar{x} - \mu)^2)\right] \end{aligned}$$

$$\text{where } S = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Therefore, the posterior is (dropping the hyperparameters as conditioning factors):

$$\begin{aligned}
p(\mu, \sigma^2 \mid \mathbf{X}) &\propto p(\mu, \sigma^2) p(\mathbf{X} \mid \mu, \sigma^2) \\
&\propto (\sigma^2)^{-(\nu_0+3)/2} \exp\left[-\frac{1}{2\sigma^2} (\nu_0\sigma_0^2 + n_0(\mu - \mu_0)^2)\right] \sigma^{2-n/2} \exp\left[-\frac{1}{2\sigma^2} (S + n(\bar{x} - \mu)^2)\right] \\
&= (\sigma^2)^{-(\nu_0+n+3)/2} \exp\left[-\frac{1}{2\sigma^2} (\nu_0\sigma_0^2 + S + n_0(\mu - \mu_0)^2 + n(\bar{x} - \mu)^2)\right] \\
&= (\sigma^2)^{-(\nu_0+n+3)/2} \exp\left[-\frac{1}{2\sigma^2} \left(\nu_0\sigma_0^2 + S + \frac{n_0n}{n_0+n}(\mu_0 - \bar{x})^2 + (n_0+n)\left(\mu - \frac{n_0\mu_0 + n\bar{x}}{n_0+n}\right)^2\right)\right] \\
&\propto (\sigma^2)^{-1/2} \exp\left[-\frac{n_0+n}{2\sigma^2} \left(\mu - \frac{n_0\mu_0 + n\bar{x}}{n_0+n}\right)^2\right] \\
&\quad \times (\sigma^2)^{-(\nu_0/2+n/2+1)} \exp\left[-\frac{1}{2\sigma^2} \left(\nu_0\sigma_0^2 + S + \frac{n_0n}{n_0+n}(\mu_0 - \bar{x})^2\right)\right] \\
&= \mathcal{N}_{\mu|\sigma^2} \left(\frac{n_0\mu_0 + n\bar{x}}{n_0+n}, \frac{\sigma^2}{n_0+n} \right) \cdot \text{IG}_{\sigma^2} \left(\frac{1}{2}(\nu_0+n), \frac{1}{2} \left(\nu_0\sigma_0^2 + S + \frac{n_0n}{n_0+n}(\mu_0 - \bar{x})^2 \right) \right).
\end{aligned}$$

In other words, the posterior distribution has the form of a product of a normal distribution over $p(\mu \mid \sigma^2)$ times an inverse gamma distribution over $p(\sigma^2)$, with parameters that are the same as the update equations above.

Occurrence and applications

The occurrence of normal distribution in practical problems can be loosely classified into four 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.
4. Regression problems – the normal distribution being found after systematic effects have been modeled sufficiently well.

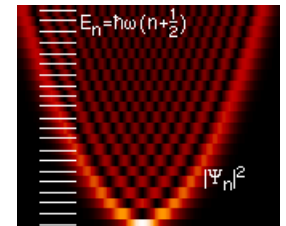
Exact normality

Certain quantities in physics are distributed normally, as was first demonstrated by James Clerk Maxwell. Examples of such quantities are:

- 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 $\frac{\partial}{\partial t} f(x, t) = \frac{1}{2} \frac{\partial^2}{\partial x^2} f(x, t)$. 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.

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.



The ground state of a quantum harmonic oscillator has the Gaussian distribution.

- 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 radiation has a Bose–Einstein distribution on very short time scales, and a normal distribution on longer timescales due to the central limit theorem.

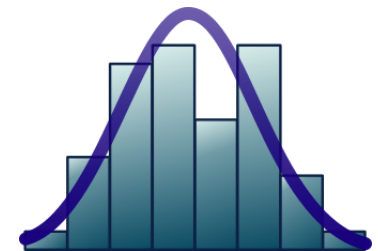
Assumed normality

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.

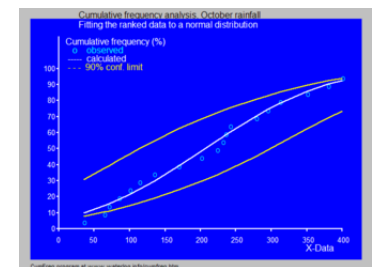
— Pearson (1901)

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);^[49]
 - 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 Benoit 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. The use of the assumption of normal distribution occurring in financial models has also been criticized by Nassim Nicholas Taleb in his works.
- 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.^[50]
- 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.^[51] The blue picture, made with CumFreq, 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.



Histogram of sepal widths for *Iris versicolor* from Fisher's Iris flower data set, with superimposed best-fitting normal distribution.



Fitted cumulative normal distribution to October rainfalls, see distribution fitting

Produced normality

In regression analysis, lack of normality in residuals simply indicates that the model postulated is inadequate in accounting for the tendency in the data and needs to be augmented; in other words, normality in residuals can always be achieved given a properly constructed model.

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,^[52] 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 $(-6, 6)$.^[53]
- 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

$$X = \sqrt{-2\ln U} \cos(2\pi V), \quad Y = \sqrt{-2\ln U} \sin(2\pi V).$$

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 .

- The Marsaglia polar method is a modification of the Box–Muller method which does not require computation of the sine and cosine functions. 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 1, then the method starts over, otherwise the 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 are independent, standard normal random variables.

- The Ratio method^[54] 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 two optional steps allow the evaluation of the logarithm in the last step to be avoided in most cases. These steps can be greatly improved^[55] so that the logarithm is rarely evaluated.

- The ziggurat algorithm^[56] 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.
- Integer arithmetic can be used to sample from the standard normal distribution.^[57] This method is exact in the sense that it satisfies the conditions of *ideal approximation*,^[58] i.e., it is equivalent to sampling a real number from the standard normal distribution and rounding this to the nearest representable floating point number.
- There is also some investigation^[59] 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.



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.

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 (http://www.math.sfu.ca/~cbm/aands/page_932.htm)):

$$\Phi(x) = 1 - \varphi(x) (b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5) + \varepsilon(x), \quad t = \frac{1}{1 + b_0 x},$$

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 some dozens of approximations – by means of rational functions, with or without exponentials – for the $\operatorname{erfc}()$ 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^[note 2] based on the Taylor series expansion

$$\Phi(x) = \frac{1}{2} + \varphi(x) \left(x + \frac{x^3}{3} + \frac{x^5}{3 \cdot 5} + \frac{x^7}{3 \cdot 5 \cdot 7} + \frac{x^9}{3 \cdot 5 \cdot 7 \cdot 9} + \dots \right)$$

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.

Shore (1982) introduced simple approximations that may be incorporated in stochastic optimization models of engineering and operations research, like reliability engineering and inventory analysis. Denoting $p = \Phi(z)$, the simplest approximation for the quantile function is:

$$z = \Phi^{-1}(p) = 5.5556 \left[1 - \left(\frac{1-p}{p} \right)^{0.1186} \right], \quad p \geq 1/2$$

This approximation delivers for z a maximum absolute error of 0.026 (for $0.5 \leq p \leq 0.9999$, corresponding to $0 \leq z \leq 3.719$). For $p < 1/2$ replace p by $1 - p$ and change sign. Another approximation, somewhat less accurate, is the single-parameter approximation:

$$z = -0.4115 \left\{ \frac{1-p}{p} + \log \left[\frac{1-p}{p} \right] - 1 \right\}, \quad p \geq 1/2$$

The latter had served to derive a simple approximation for the loss integral of the normal distribution, defined by

$$L(z) = \int_z^{\infty} (u - z)\varphi(u) \, du = \int_z^{\infty} [1 - \Phi(u)] \, du$$

$$L(z) \approx \begin{cases} 0.4115 \left(\frac{p}{1-p} \right) - z, & p < 1/2, \\ 0.4115 \left(\frac{1-p}{p} \right), & p \geq 1/2. \end{cases}$$

or, equivalently,

$$L(z) \approx \begin{cases} 0.4115 \left\{ 1 - \log \left[\frac{p}{1-p} \right] \right\}, & p < 1/2, \\ 0.4115 \frac{1-p}{p}, & p \geq 1/2. \end{cases}$$

This approximation is particularly accurate for the right far-tail (maximum error of 10^{-3} for $z \geq 1.4$). Highly accurate approximations for the CDF, based on Response Modeling Methodology (RMM, Shore, 2011, 2012), are shown in Shore (2005).

Some more approximations can be found at: Error function#Approximation with elementary functions. In particular, small *relative* error on the whole domain for the CDF Φ and the quantile function Φ^{-1} as well, is achieved via an explicitly invertible formula by Sergei Winitzki in 2008.

History

Development

Some authors^{[60][61]} attribute the credit for the discovery of the normal distribution to de Moivre, who in 1738^[note 3] 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 ℓ , has to the middle Term, is $-\frac{2\ell\ell}{n}$ ".^[62] 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.^[63]

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 of that quantity: 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.^[note 4] 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.^[64]

$$\varphi\Delta = \frac{h}{\sqrt{\pi}} e^{-h^2\Delta\Delta},$$

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.^[65]



Carl Friedrich Gauss discovered the normal distribution in 1809 as a way to rationalize the method of least squares.

Although Gauss was the first to suggest the normal distribution law, Laplace made significant contributions.^[note 5] It was Laplace who first posed the problem of aggregating several observations in 1774,^[66] 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.^[67] 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.^[68]

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.^[69] His works remained largely unnoticed by the scientific community, until in 1871 they were "rediscovered" by Abbe.^[70]

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:^[71] "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".^[72] However, by the end of the 19th century some authors^[note 6] 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."^[73] Around the turn of the 20th century Pearson popularized the term *normal* as a designation for this distribution.^[74]

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:

$$df = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-(x-m)^2/(2\sigma^2)} dx$$

The term "standard normal", which denotes the normal distribution with zero mean and unit variance came into general use around the 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*".^[75]

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 speakers, both "normal distribution" and "Gaussian distribution" are in common use, with different terms preferred by different communities.

See also

- Wrapped normal distribution — the Normal distribution applied to a circular domain



Marquis de Laplace proved the central limit theorem in 1810, consolidating the importance of the normal distribution in statistics.

- Bates distribution — similar to the Irwin–Hall distribution, but rescaled back into the 0 to 1 range
- Behrens–Fisher problem — the long-standing problem of testing whether two normal samples with different variances have same means;
- Bhattacharyya distance – method used to separate mixtures of normal distributions
- Erdős–Kac theorem—on the occurrence of the normal distribution in number theory
- Gaussian blur—convolution, which uses the normal distribution as a kernel
- Normally distributed and uncorrelated does not imply independent
- Standard normal table
- Sub-Gaussian distribution
- Sum of normally distributed random variables
- Tweedie distribution — The normal distribution is a member of the family of Tweedie exponential dispersion models
- Z-test— using the normal distribution
- Stein's lemma

Notes

1. For the proof see **Gaussian integral**.
2. For example, this algorithm is given in the article **Bc programming language**.
3. De Moivre first published his findings in 1733, in a pamphlet "Approximatio ad Summam Terminorum Binomii $(a + b)^n$ 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 **Walker (1985)**.
4. "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." — **Gauss (1809**, section 177)
5. "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 **Pearson (1905**, p. 189)
6. Besides those specifically referenced here, such use is encountered in the works of **Peirce**, **Galton** (**Galton (1889**, chapter V)) and **Lexis** (**Lexis (1878)**, **Rohrbasser & Véron (2003)**) c. 1875.

References

Citations

1. *Normal Distribution* (http://www.encyclopedia.com/topic/Normal_Distribution.aspx#3), Gale Encyclopedia of Psychology
2. **Casella & Berger** (2001, p. 102)
3. **Lyon, A.** (2014). Why are Normal Distributions Normal? (https://aidanlyon.com/normal_distributions.pdf), The British Journal for the Philosophy of Science.
4. **Stigler** (1982)
5. **Halperin, Hartley & Hoel** (1965, item 7)
6. **McPherson** (1990, p. 110)
7. **Bernardo & Smith** (2000, p. 121)
8. **Cover, Thomas M.; Thomas, Joy A.** (2006). *Elements of Information Theory*. John Wiley and Sons. p. 254.
9. **Park, Sung Y.; Bera, Anil K.** (2009). "Maximum Entropy Autoregressive Conditional Heteroskedasticity Model" (<http://www.wise.xmu.edu.cn/Master/Download/..%5C..%5CUploadFiles%5Cpaper-masterdownload%5C2009519932327055475115776.pdf>) (PDF). *Journal of Econometrics*. **150** (2): 219–230. CiteSeerX 10.1.1.511.9750 (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.511.9750>). doi:10.1016/j.jeconom.2008.12.014 (<https://doi.org/10.1016%2Fj.jeconom.2008.12.014>). Retrieved June 2, 2011.
10. **Geary RC**(1936) The distribution of the "Student's" ratio for the non-normal samples". Supplement to the Journal of the Royal Statistical Society 3 (2): 178–184
11. **Lukas E** (1942) A characterization of the normal distribution. Annals of Mathematical Statistics 13: 91–93
12. **Patel & Read** (1996, [2.1.4])
13. **Fan** (1991, p. 1258)
14. **Patel & Read** (1996, [2.1.8])
15. **Papoulis, Athanasios.** *Probability, Random Variables and Stochastic Processes (4th Edition)*. p. 148.
16. **Bryc** (1995, p. 23)
17. **Bryc** (1995, p. 24)

18. Scott, Clayton; Nowak, Robert (August 7, 2003). "The Q-function" (<http://cnx.org/content/m11537/1.2/>). *Connexions*.
19. Barak, Ohad (April 6, 2006). "Q Function and Error Function" (<https://web.archive.org/web/20090325160012/http://www.eng.tau.ac.il/~jo/academic/Q.pdf>) (PDF). Tel Aviv University. Archived from the original (<http://www.eng.tau.ac.il/~jo/academic/Q.pdf>) (PDF) on March 25, 2009.
20. Weisstein, Eric W. "Normal Distribution Function" (<http://mathworld.wolfram.com/NormalDistributionFunction.html>). *MathWorld*.
21. Abramowitz, Milton; Stegun, Irene Ann, eds. (1983) [June 1964]. "Chapter 26, eqn 26.2.12" (http://www.math.sfu.ca/~cbm/aands/page_932.htm). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Applied Mathematics Series. **55** (Ninth reprint with additional corrections of tenth original printing with corrections (December 1972); first ed.). Washington D.C.; New York: United States Department of Commerce, National Bureau of Standards; Dover Publications. p. 932. ISBN 978-0-486-61272-0. LCCN 64-60036 (<https://lccn.loc.gov/64-60036>). MR 0167642 (<https://www.ams.org/mathscinet-getitem?mr=0167642>). LCCN 65-12253 (<https://lccn.loc.gov/65012253>).
22. "Wolfram|Alpha: Computational Knowledge Engine" ([http://www.wolframalpha.com/input/?i=Table%5B{N\(Erf\(n/Sqrt\(2\)\)\),+12},+N\(1-Erf\(n/Sqrt\(2\)\)\),+12},+N\(1/\(1-Erf\(n/Sqrt\(2\)\)\),+12\)},+{n,1,6}%5D](http://www.wolframalpha.com/input/?i=Table%5B{N(Erf(n/Sqrt(2))),+12},+N(1-Erf(n/Sqrt(2))),+12},+N(1/(1-Erf(n/Sqrt(2))),+12)},+{n,1,6}%5D)). *Wolframalpha.com*. Retrieved March 3, 2017.
23. "Wolfram|Alpha: Computational Knowledge Engine" (http://www.wolframalpha.com/input/?i=Table%5B5Bqrt%282%29*InverseErf%28x%29%2C+{x%2C+N%288%2F10%2C+9%2F10%2C+19%2F20%2C+49%2F50%2C+99%2F100%2C+995%2F1000%2C+998%2F1000}%2C+13%29}%5D). *Wolframalpha.com*.
24. "Wolfram|Alpha: Computational Knowledge Engine" ([http://www.wolframalpha.com/input/?i=Table%5B%7BN\(1-10%5E\(-x\),9\),N\(Sqrt\(2\)*InverseErf\(1-10%5E\(-x\)\),13\)%7D,%7Bx,3,9%7D}%5D](http://www.wolframalpha.com/input/?i=Table%5B%7BN(1-10%5E(-x),9),N(Sqrt(2)*InverseErf(1-10%5E(-x)),13)%7D,%7Bx,3,9%7D}%5D)). *Wolframalpha.com*. Retrieved March 3, 2017.
25. "Normal Approximation to Poisson Distribution" (http://www.stat.ucla.edu/~dinov/courses_students.dir/Applets.dir/NormalApprox2PoissonApplet.html). *Stat.ucla.edu*. Retrieved March 3, 2017.
26. Cover & Thomas (2006, p. 254)
27. Williams, David (2001). *Weighing the odds : a course in probability and statistics* (Reprinted. ed.). Cambridge [u.a.]: Cambridge Univ. Press. pp. 197–199. ISBN 978-0-521-00618-7.
28. Smith, José M. Bernardo; Adrian F. M. (2000). *Bayesian theory* (Reprint ed.). Chichester [u.a.]: Wiley. pp. 209, 366. ISBN 978-0-471-49464-5.
29. O'Hagan, A. (1994) *Kendall's Advanced Theory of statistics, Vol 2B, Bayesian Inference*, Edward Arnold. ISBN 0-340-52922-9 (Section 5.40)
30. Bryc (1995, p. 27)
31. Patel & Read (1996, [2.3.6])
32. Galambos & Simonelli (2004, Theorem 3.5)
33. Bryc (1995, p. 35)
34. Lukacs & King (1954)
35. 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.
36. 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".
37. Edward L. Melnick and Aaron Tenenbein, "Misspecifications of the Normal Distribution", *The American Statistician*, volume 36, number 4 November 1982, pages 372–373
38. "Kullback Leibler (KL) Distance of Two Normal (Gaussian) Probability Distributions" (<http://www.allisons.org/ll/MML/KL/Normal/>). *Allisons.org*. December 5, 2007. Retrieved March 3, 2017.
39. Jordan, Michael I. (February 8, 2010). "Stat260: Bayesian Modeling and Inference: The Conjugate Prior for the Normal Distribution" (<http://www.cs.berkeley.edu/~jordan/courses/260-spring10/lectures/lecture5.pdf>) (PDF).
40. Amari & Nagaoka (2000)
41. Weisstein, Eric W. "Normal Product Distribution" (<http://mathworld.wolfram.com/NormalProductDistribution.html>). *MathWorld*. wolfram.com.
42. Lukacs, Eugene (1942). "A Characterization of the Normal Distribution". *The Annals of Mathematical Statistics*. **13** (1): 91–3. doi:10.1214/aoms/1177731647 (<https://doi.org/10.1214%2Faoms%2F1177731647>). ISSN 0003-4851 (<https://www.worldcat.org/issn/0003-4851>). JSTOR 2236166 (<https://www.jstor.org/stable/2236166>).
43. Basu, D.; Laha, R. G. (1954). "On Some Characterizations of the Normal Distribution". *Sankhyā*. **13** (4): 359–62. ISSN 0036-4452 (<https://www.worldcat.org/issn/0036-4452>). JSTOR 25048183 (<https://www.jstor.org/stable/25048183>).
44. Lehmann, E. L. (1997). *Testing Statistical Hypotheses* (2nd ed.). Springer. p. 199. ISBN 978-0-387-94919-2.
45. John, S (1982). "The three parameter two-piece normal family of distributions and its fitting". *Communications in Statistics - Theory and Methods*. **11** (8): 879–885. doi:10.1080/03610928208828279 (<https://doi.org/10.1080%2F03610928208828279>).
46. Krishnamoorthy (2006, p. 127)
47. Krishnamoorthy (2006, p. 130)
48. Krishnamoorthy (2006, p. 133)
49. Huxley (1932)
50. Jaynes, Edwin T. (2003). *Probability Theory: The Logic of Science* (<https://books.google.com/books?id=tTN4HuUNXjgC&pg=PA592>). Cambridge University Press. pp. 592–593. ISBN 9780521592710.
51. Oosterbaan, Roland J. (1994). "Chapter 6: Frequency and Regression Analysis of Hydrologic Data" (<http://www.waterlog.info/pdf/freqtxt.pdf>) (PDF). In Ritzema, Henk P. (ed.). *Drainage Principles and Applications, Publication 16* (second revised ed.). Wageningen, The Netherlands: International Institute for Land Reclamation and Improvement (ILRI). pp. 175–224. ISBN 978-90-70754-33-4.
52. Wichura, Michael J. (1988). "Algorithm AS241: The Percentage Points of the Normal Distribution". *Applied Statistics*. **37** (3): 477–84. doi:10.2307/2347330 (<https://doi.org/10.2307%2F2347330>). JSTOR 2347330 (<https://www.jstor.org/stable/2347330>).
53. Johnson, Kotz & Balakrishnan (1995, Equation (26.48))
54. Kinderman & Monahan (1977)
55. Leva (1992)
56. Marsaglia & Tsang (2000)
57. Karney (2016)
58. Monahan (1985, section 2)
59. Wallace (1996)
60. Johnson, Kotz & Balakrishnan (1994, p. 85)
61. Le Cam & Lo Yang (2000, p. 74)
62. De Moivre, Abraham (1733), Corollary I – see Walker (1985, p. 77)
63. Stigler (1986, p. 76)

- 64. Gauss (1809, section 177)
- 65. Gauss (1809, section 179)
- 66. Laplace (1774, Problem III)
- 67. Pearson (1905, p. 189)
- 68. Stigler (1986, p. 144)
- 69. Stigler (1978, p. 243)

- 70. Stigler (1978, p. 244)
- 71. Maxwell (1860, p. 23)
- 72. Jaynes, Edwin J.; *Probability Theory: The Logic of Science*, Ch 7 (<http://www.biba.inrialpes.fr/Jaynes/c07s.pdf>)
- 73. Peirce, Charles S. (c. 1909 MS), *Collected Papers* v. 6, paragraph 327
- 74. Kruskal & Stigler (1997)
- 75. "Earliest uses... (entry STANDARD NORMAL CURVE)" (<http://jeff560.tripod.com/s.html>).

Sources

- 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 978-0-8218-0531-2.
- Bernardo, José M.; Smith, Adrian F. M. (2000). *Bayesian Theory*. Wiley. ISBN 978-0-471-49464-5.
- Bryc, Włodzimierz (1995). *The Normal Distribution: Characterizations with Applications*. Springer-Verlag. ISBN 978-0-387-97990-8.
- Casella, George; Berger, Roger L. (2001). *Statistical Inference* (2nd ed.). Duxbury. ISBN 978-0-534-24312-8.
- 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 (<https://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 978-0-8218-2103-9.
- 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 (<https://doi.org/10.1214/aos/1176348248>) (JSTOR 2241949 (<https://www.jstor.org/stable/2241949>)).
- Galton, Francis (1889). *Natural Inheritance* (<http://galton.org/books/natural-inheritance/pdf/galton-nat-inh-1up-clean.pdf>) (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 978-0-8247-5402-0.
- 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 (<https://books.google.com/books?id=1TIAAAAAQAAJ>).
- Gould, Stephen Jay (1981). *The Mismeasure of Man* (first ed.). W. W. Norton. ISBN 978-0-393-01489-1.
- 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 (<https://doi.org/10.2307/2681417>). JSTOR 2681417 (<https://www.jstor.org/stable/2681417>).
- Hart, John F.; et al. (1968). *Computer Approximations*. New York, NY: John Wiley & Sons, Inc. ISBN 978-0-88275-642-4.
- Hazewinkel, Michiel, ed. (2001) [1994], "Normal Distribution" (<https://www.encyclopediaofmath.org/index.php?title=p/n067460>), *Encyclopedia of Mathematics*, Springer Science+Business Media B.V. / Kluwer Academic Publishers, 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 978-0-02-914673-6.
- Huxley, Julian S. (1932). *Problems of Relative Growth*. London. ISBN 978-0-486-61114-3. OCLC 476909537 (<https://www.worldcat.org/oclc/476909537>).
- Johnson, Norman L.; Kotz, Samuel; Balakrishnan, Narayanaswamy (1994). *Continuous Univariate Distributions, Volume 1*. Wiley. ISBN 978-0-471-58495-7.
- Johnson, Norman L.; Kotz, Samuel; Balakrishnan, Narayanaswamy (1995). *Continuous Univariate Distributions, Volume 2*. Wiley. ISBN 978-0-471-58494-0.
- Karney, C. F. F. (2016). "Sampling exactly from the normal distribution". *ACM Transactions on Mathematical Software*. **42** (1): 3:1–14. arXiv:1303.6257 (<https://arxiv.org/abs/1303.6257>). doi:10.1145/2710016 (<https://doi.org/10.1145/2710016>).
- 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 (<https://doi.org/10.1145/355744.355750>).
- Krishnamoorthy, Kalimuthu (2006). *Handbook of Statistical Distributions with Applications*. Chapman & Hall/CRC. ISBN 978-1-58488-635-8.

- 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 978-0-19-852341-3.
- 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 (<https://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 978-0-387-95036-5.
- Leva, Joseph L. (1992). "A fast normal random number generator" (<https://web.archive.org/web/20100716035328/http://saluc.engr.uconn.edu/refs/crypto/rng/leva92afast.pdf>) (PDF). *ACM Transactions on Mathematical Software*. 18 (4): 449–453. CiteSeerX 10.1.1.544.5806 (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.544.5806>). doi:10.1145/138351.138364 (<https://doi.org/10.1145/138351.138364>). Archived from the original (<http://saluc.engr.uconn.edu/refs/crypto/rng/leva92afast.pdf>) (PDF) on July 16, 2010.
- 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 (<https://doi.org/10.1214/aoms/1177728796>). JSTOR 2236741 (<https://www.jstor.org/stable/2236741>).
- McPherson, Glen (1990). *Statistics in Scientific Investigation: Its Basis, Application and Interpretation*. Springer-Verlag. ISBN 978-0-387-97137-7.
- Marsaglia, George; Tsang, Wai Wan (2000). "The Ziggurat Method for Generating Random Variables". *Journal of Statistical Software*. 5 (8). doi:10.18637/jss.v005.i08 (<https://doi.org/10.18637/jss.v005.i08>).
- Marsaglia, George (2004). "Evaluating the Normal Distribution". *Journal of Statistical Software*. 11 (4). doi:10.18637/jss.v011.i04 (<https://doi.org/10.18637/jss.v011.i04>).
- 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 (<https://doi.org/10.1080/14786446008642818>).
- Monahan, J. F. (1985). "Accuracy in random number generation". *Mathematics of Computation*. 45 (172): 559–568. doi:10.1090/S0025-5718-1985-0804945-X (<https://doi.org/10.1090/S0025-5718-1985-0804945-X>).
- Patel, Jagdish K.; Read, Campbell B. (1996). *Handbook of the Normal Distribution* (2nd ed.). CRC Press. ISBN 978-0-8247-9342-5.
- Pearson, Karl (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space" (<http://stat.smmu.edu.cn/history/pearson1901.pdf>) (PDF). *Philosophical Magazine*. 6. 2 (11): 559–572. doi:10.1080/14786440109462720 (<https://doi.org/10.1080/14786440109462720>).
- Pearson, Karl (1905). "'Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson'. A rejoinder". *Biometrika*. 4 (1): 169–212. doi:10.2307/2331536 (<https://doi.org/10.2307/2331536>). JSTOR 2331536 (<https://www.jstor.org/stable/2331536>).
- Pearson, Karl (1920). "Notes on the History of Correlation" (<https://zenodo.org/record/1431597/files/article.pdf>) (PDF). *Biometrika*. 13 (1): 25–45. doi:10.1093/biomet/13.1.25 (<https://doi.org/10.1093/biomet/13.1.25>). JSTOR 2331722 (<https://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. doi:10.3917/pope.303.0303 (<https://doi.org/10.3917/pope.303.0303>).
- Shore, H (1982). "Simple Approximations for the Inverse Cumulative Function, the Density Function and the Loss Integral of the Normal Distribution". *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. 31 (2): 108–114. doi:10.2307/2347972 (<https://doi.org/10.2307/2347972>). JSTOR 2347972 (<https://www.jstor.org/stable/2347972>).
- Shore, H (2005). "Accurate RMM-Based Approximations for the CDF of the Normal Distribution". *Communications in Statistics – Theory and Methods*. 34 (3): 507–513. doi:10.1081/sta-200052102 (<https://doi.org/10.1081/sta-200052102>).
- Shore, H (2011). "Response Modeling Methodology". *WIREs Comput Stat*. 3 (4): 357–372. doi:10.1002/wics.151 (<https://doi.org/10.1002/wics.151>).
- Shore, H (2012). "Estimating Response Modeling Methodology Models". *WIREs Comput Stat*. 4 (3): 323–333. doi:10.1002/wics.1199 (<https://doi.org/10.1002/wics.1199>).
- Stigler, Stephen M. (1978). "Mathematical Statistics in the Early States". *The Annals of Statistics*. 6 (2): 239–265. doi:10.1214/aos/1176344123 (<https://doi.org/10.1214/aos/1176344123>). JSTOR 2958876 (<https://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 (<https://doi.org/10.2307/2684031>). JSTOR 2684031 (<https://www.jstor.org/stable/2684031>).
- Stigler, Stephen M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press. ISBN 978-0-674-40340-6.
- Stigler, Stephen M. (1999). *Statistics on the Table*. Harvard University Press. ISBN 978-0-674-83601-3.
- Walker, Helen M. (1985). "De Moivre on the Law of Normal Probability" (<http://www.york.ac.uk/depts/maths/histstat/demoivre.pdf>) (PDF). In Smith, David Eugene (ed.). *A Source Book in Mathematics*. Dover. ISBN 978-0-486-64690-9.
- 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 (<https://doi.org/10.1145/225545.225554>).

- 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) (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 978-0-486-61272-0.

External links

- Hazewinkel, Michiel, ed. (2001) [1994], "Normal distribution" (<https://www.encyclopediaofmath.org/index.php?title=p/n067460>), *Encyclopedia of Mathematics*, Springer Science+Business Media B.V. / Kluwer Academic Publishers, ISBN 978-1-55608-010-4
 - Normal distribution calculator (<https://www.hackmath.net/en/calculator/normal-distribution>)
-

Retrieved from "https://en.wikipedia.org/w/index.php?title=Normal_distribution&oldid=922730255"

This page was last edited on 23 October 2019, at 22:55 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.