Discrete and continuous random variables

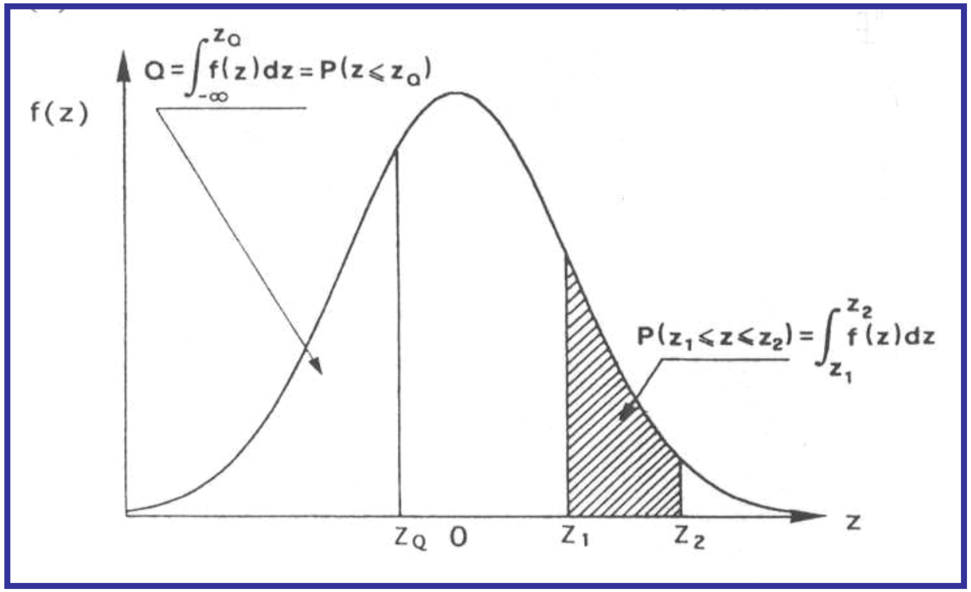
A random variable (also called aleatory or stochastic) is a variable that can assume different values, according to an aleatory phenomenon.

A typical example is the result of dice casting, which can assume only one of the possibile integer numbers comprised between 1 and 6, each with the same probability (1/6). The score of dice casting is an example of discrete random variable, i.e., a variable that can assume only values included in a subset of the set of real numbers.
Interestingly, in the case of a discrete variable no value can have a zero probability (after each dice casting a value among possible ones is certainly obtained).

A continuous random variable can assume, at least potentially, all values included in the set of real numbers.
As an example, the length of a leaf could correspond to one of the infinite values comprised between 3 and 4 cm. Speaking strictly in mathematical terms, each of the possible values obtained after measuring the leaf length has a zero probability, since its occurrence has to be divided by infinity. This is a paradox typical of continuous variables. Only the probability for such a variable to assume a value included in an interval can be calculated.

Probability density function

In probability theory, a Probability Density Function (PDF), or density of a continuous random variable z, indicated as f(z), is a function that describes the relative likelihood for this random variable to assume a given value.

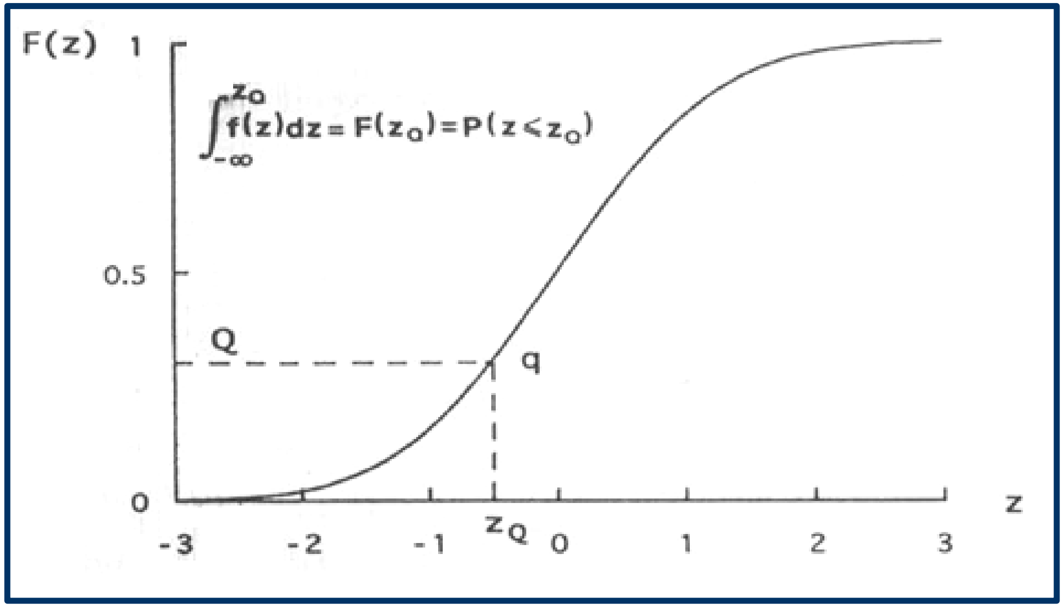


The probability of the random variable falling within a particular range of values is given by the integral of this variable's density over that range, that is, the area underlying the density function between the lowest and the greatest values of the range.
This definition clarifies that the probability referred to a single value of the variable is zero, since it would correspond to the integral calculated from z_1 to z_1.

In the picture, the probability of the random variable falling between $-\infty$ and a specific value z_Q, that is called *quantile*, is also shown.

Cumulative distribution function

In probability theory, the cumulative distribution function (CDF) of a random variable z, F(z) is a function describing the probability that the variable will assume a value less than or equal to a specific number z_Q.



Functions f(z) and F(z) are related by the following equations:

$$f(z) = dF(z)/dz$$

$$F(z_Q) = \int_{-\infty}^{z_Q} f(z)dz$$

Expectation (expected value) and variance of a continuous random variable

The expectation of a continuous random variable x whose probability density function is f(x) can be calculated as:

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)dx = \mu$$

This is a special case of the general equation:

$$E\{g(X)\} = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

expressing the expected value for a function g of the continuous random variable x.

Note that the expectation for a constant *a* (i.e., when g(X) = *a*) is, obviously, *a* itself.

When $g(X) = X^n$ the resulting expectation is defined as non-central (or simple) moment of order n (also called the n^{th} non-central moment) for the probability distribution of the random variable X. It is indicated as μ_n or m_n.

Variance of a continuous random variable X, representing a special case of central moment for a random variable (it corresponds to the central moment of order 2) can be calculated as:

$$V(X) = E[(X\text{-}E(X))^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx = \sigma^2$$

The following equations can be easily inferred:

$$V(x) = E(x-\mu)^2 = E(x^2 - 2x\mu + \mu^2) = E(x^2) - 2\,\mu E(x) + \mu^2 = E(x^2) - \cancel{2}\mu^2 + \cancel{\mu}^2$$

thus:

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

A consequence of this relation is that the variance of a constant *a* is 0:

$$V(a) = E(a^2)\text{-}[E(a)]^2 = a^2 - [a]^2 = 0$$

If the variable ax + b is considered, with a and b corresponding to constants, its variance can be calculated by applying previous relations:

$$V(ax + b) = V(ax) + V(b) = V(ax) = E\{(ax - a\mu)^2\} =$$

$$= E(ax)^2 + E(a\mu)^2 - 2a^2 \mu\, E(x) =$$

$$= a^2 E(x^2) + a^2 E(\mu^2) - 2a^2\mu^2 =$$

$$= a^2 E(x^2) + a^2\mu^2 - 2a^2\mu^2 =$$

$$= a^2 E(x^2) - a^2\mu^2 = a^2 [E(x^2) - \mu^2] = a^2 V(x)$$

Central moments of a continuous random variable

Variance is just one of the possible central moments of the probability distribution of a random variable.

Indeed, as a general definition, a central moment is the expected value of a specified integer power of the deviation of the random variable from the mean:

$$\mu'_n = \mathrm{E}[(X - \mathrm{E}[X])^n] = \int_{-\infty}^{+\infty} (x-\mu)^n f(x)\,\mathrm{d}x.$$

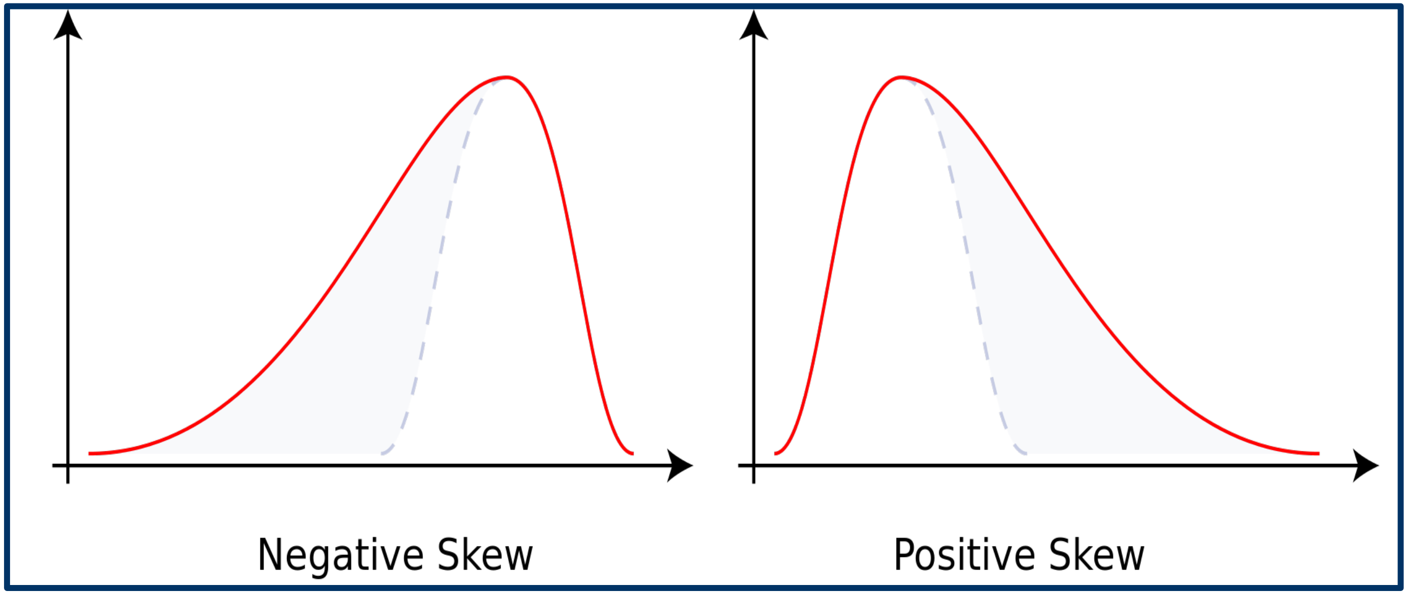
Note that an apex is often added to the μ_n symbol to distinguish central from non-central moments.

Other interesting central moments, with $n > 2$, can be calculated for the probability distribution of a random variable.

Skewness of a distribution

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

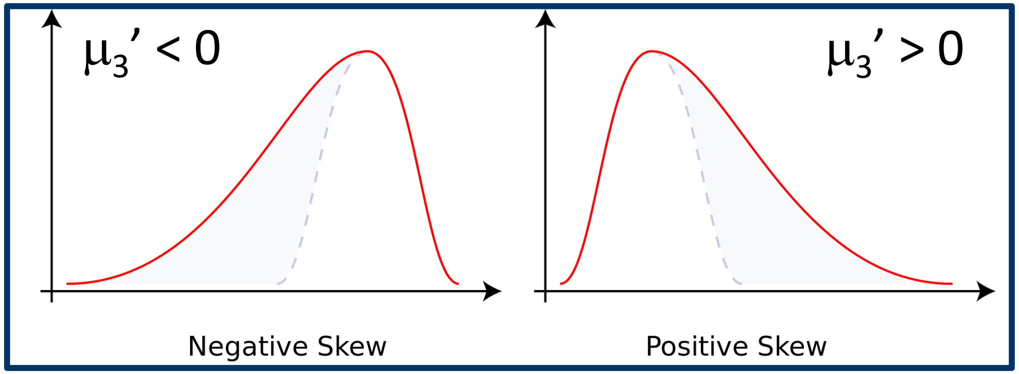
For a unimodal distribution, negative skew commonly indicates that the tail is on the left side of the distribution, and positive skew indicates that the tail is on the right:



From a mathematical point of view skewness can be calculated using the third central moment:

$$\mu_3' = E[(X-\mu)^3]$$

Negative or positive skewness correspond to the sign of this moment:



Note that $\mu_3' = 0$ in the case of a symmetric probability density function.

The same occurs for all μ_n' values with n being an odd number.

A further mathematical definition of skewness, indicated by γ_1, also called asymmetry coefficient, was given by Karl Pearson as the standardized third central moment, i.e., the third central moment ratioed to the cubic power of standard deviation:

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

Kurtosis of a distribution

Kurtosis (from Greek: κυρτός, *kyrtos* or *kurtos*, meaning "curved, arching") is a measure of the "tailedness" of the probability distribution of a real-valued random variable.
In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution and there are different ways of quantifying it for a theoretical distribution.

The standard measure of kurtosis, also proposed by Karl Pearson, correspond to the standardized fourth central moment of the population, i.e., the fourth central moment ratioed to the fourth power of standard deviation:

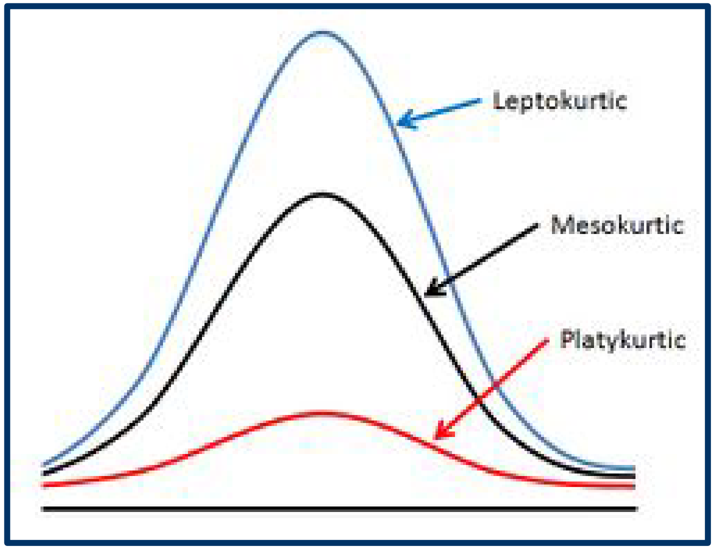
$$\mathrm{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4'}{\sigma^4} = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2}$$

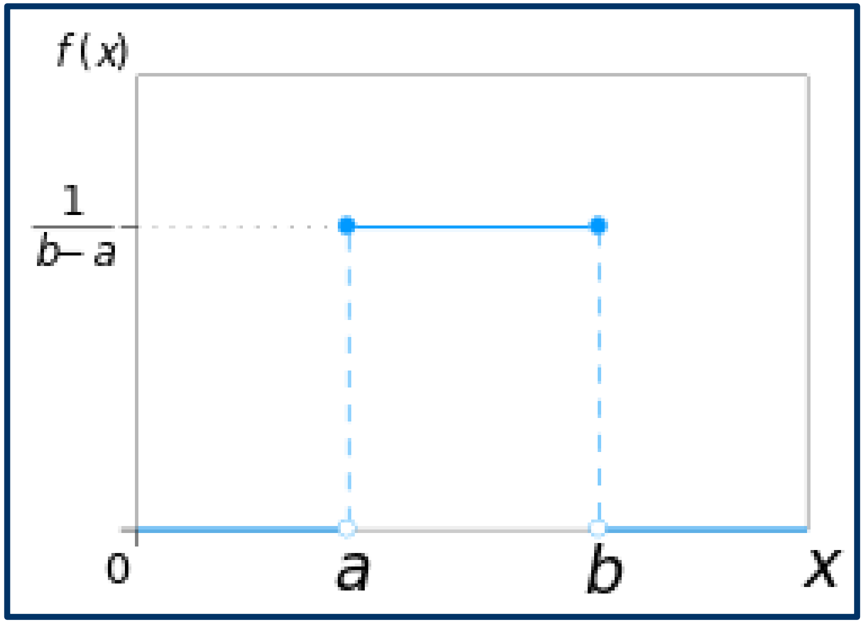
The kurtosis of any Gaussian distribution is 3 and it is common to compare the kurtosis of a distribution to this value. By general definition, distributions that have a kurtosis equal to 3 are defined “mesokurtic”.

Sometimes the difference between kurtosis and 3, called *excess kurtosis*, is defined.

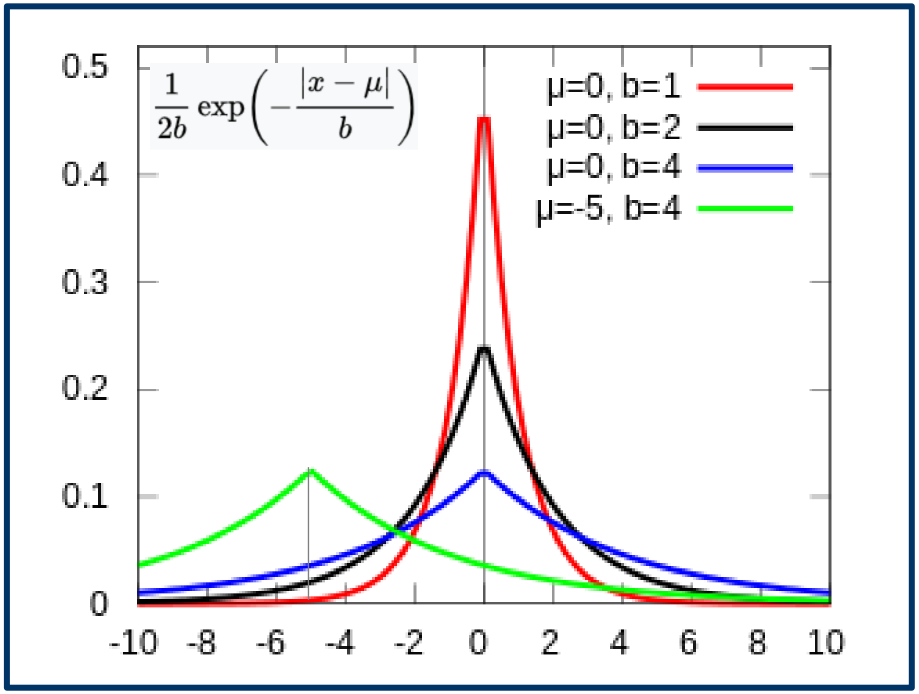
Distributions with kurtosis less than 3 are defined *platykurtic*; an extreme example is represented by the uniform distribution.

Distributions with kurtosis higher than 3 are defined leptokurtic; an example is represented by the Laplace distribution.





Uniform distribution PDF



Laplace distribution PDF

Entropy of a distribution

In statistics, the entropy of a continuous random variable having a PDF given by f(x) correspond to the expectation of the negative natural logarithm of the PDF:

$$h[f] = \mathrm{E}[-\ln(f(x))] = -\int_{\mathbb{X}} f(x)\ln(f(x))\,dx.$$

This mathematical equation is the extension to a continuous random variable of the information entropy introduced in 1948 by Claude Shannon for a discrete random variable:

$$S = -\sum_i P_i \log P_i$$

where P_i represents the probability associated to the i^{th} value of the discrete variable.

The comparison between coin toss and die casting clarifies the practical significance of information entropy.

For a fair coin toss $P_i = 1/2$ both for “heads” and for “tails”, thus $S = \log 2$. For a fair die casting $P_i = 1/6$ for each of the six possible scores (1,2,…, 6) and $S = \log 6$.

Information entropy is clearly higher in the case of die casting, since it is more difficult to predict which event (which score, for a die) will occur, compared to coin toss.

Moment-generating function (MGF)

In probability theory and statistics, the moment-generating function of a real-valued random variable is an alternative specification of its probability distribution. It enables an easy calculation of non-central (simple) moments related to a distribution.

By definition, the MGF of a random variable X is given by:

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx$$

Since the series expansion of e^{tX} is:

$$e^{t\,X} = 1 + t\,X + \frac{t^2\,X^2}{2!} + \frac{t^3\,X^3}{3!} + \cdots + \frac{t^n\,X^n}{n!} + \cdots$$

$M_X(t)$ can be also expressed as:

$$\begin{aligned} M_X(t) = \mathrm{E}(e^{t\,X}) &= 1 + t\,\mathrm{E}(X) + \frac{t^2\,\mathrm{E}(X^2)}{2!} + \frac{t^3\,\mathrm{E}(X^3)}{3!} + \cdots + \frac{t^n\,\mathrm{E}(X^n)}{n!} + \cdots \\ &= 1 + tm_1 + \frac{t^2 m_2}{2!} + \frac{t^3 m_3}{3!} + \cdots + \frac{t^n m_n}{n!} + \cdots, \end{aligned}$$

where m_n are different non-central (simple) moments of the distribution.

By definition, the n^th^ moment of a probability density function, indicated by m_n, is the expectation of the n^th^ power of the corresponding random variable:

$$m_n = \mathrm{E}[X^n] = \int_{-\infty}^{\infty} x^n f(x)\, \mathrm{d}x$$

The most used moment of a PDF is the first (n = 1), corresponding to the population mean.

If $M_X(t)$ is differentiated k times with respect to t and t is set equal to 0 the k^th^ non-central moment of the distribution is obtained.

In fact:

$$[\mathrm{d}M_X(t)/\mathrm{d}t] = m_1 + \frac{2\,t\ m_2}{2!} + \frac{3\,t^2 m_3}{3!} + \cdots$$

thus:

$$[\mathrm{d}M_X(t)/\mathrm{d}t]_{t=0} = E(X) = m_1 \qquad [\mathrm{d}^2M_X(t)/\mathrm{d}t^2]_{t=0} = E(X^2) = m_2$$

and so on.

Median and mode of a distribution

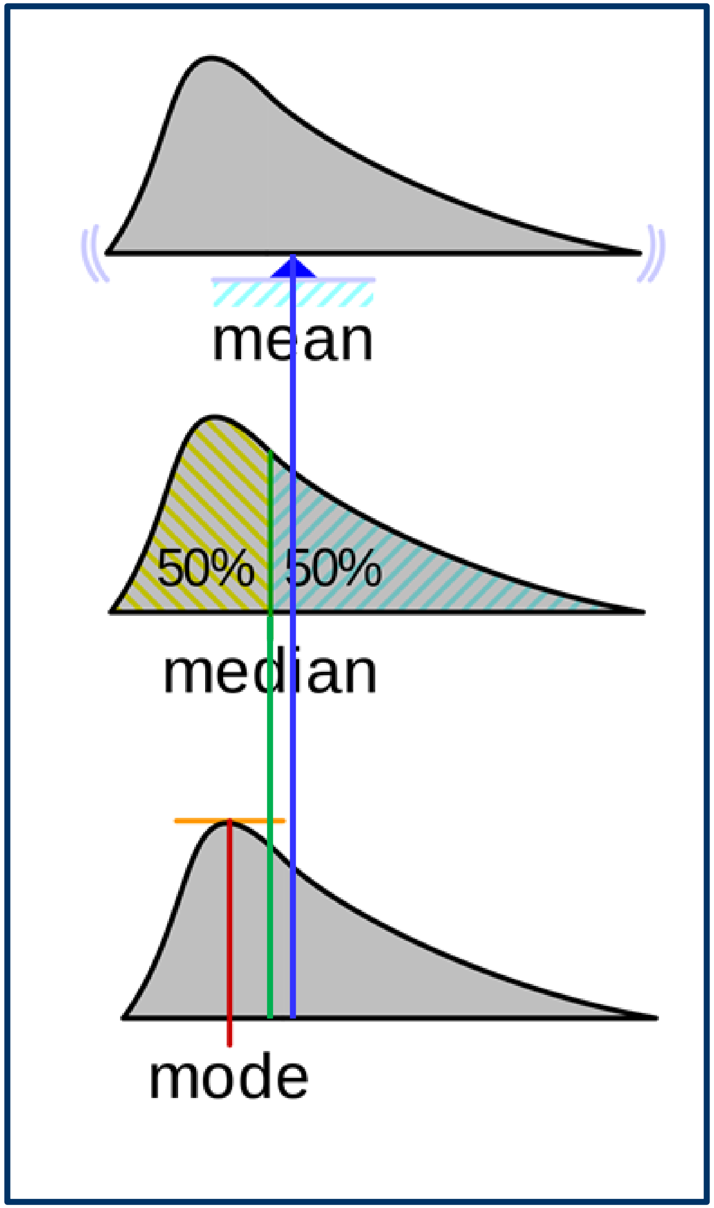
Median and mode can be used as alternatives to the mean for the description of location (central tendency) of a distribution on the axis reporting the variable values.

From a mathematical point of view, the median, indicated as m, can be inferred from the equations:

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

Mode is the value corresponding to the maximum of the f(x) function (note that more than one maximum can be present in a multimodal distribution).

Mean, mode and median do not coincide in the case of unimodal, asymmetric distributions (median is located between mode and mean).



Covariance

Given two random variables X and Y, with mean μ_X and μ_y, respectively, their covariance, indicated as cov[X,Y] or as $\sigma_{X,Y}$, is given by the equation:

$$\mathrm{cov}[X,Y] = E[(X - \mu_X)(Y - \mu_y)]$$

Some relevant properties of covariance are:

- $\mathrm{cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y];$
- $\mathrm{cov}[X, Y] = \mathrm{cov}[Y, X];$
- $\mathrm{cov}[aX, bY] = ab\ \mathrm{cov}[X, Y];$
- $\mathrm{cov}[X, X] = \sigma_X^2;$
- $\mathrm{cov}[X + Y, Z] = \mathrm{cov}[X, Z] + \mathrm{cov}[Y, Z].$

The first property can be demonstrated by considering some properties of Expectation:

$$\text{cov}[X,Y] = E[(X-\mu_x)(Y-\mu_y)] = E[XY - \mu_y X - \mu_x Y + \mu_x\mu_y] =$$

$$= E(XY) - \mu_y E(X) - \mu_x E(Y) + \mu_x\mu_y =$$

$$= E(XY) - \mu_y\mu_x - \mu_x\mu_y + \mu_x\mu_y =$$

$$= E(XY) - \mu_x\mu_y = E(XY) - E(X)E(Y)$$

Note that, if Y = X, the final equation becomes:

$$V(X) = E(X^2)-[E(X)]^2 = E(X^2)-\mu^2$$

thus:

$$\text{cov}[X,X] = V(X)$$

If random variables X and Y are independent, i.e., if their probabilities do not influence each other, for any couple of set of values (A,B) the following equation is true:

$$P(X \in A, Y \in B) = P(X \in A) * P(Y \in B)$$

In this case:

$$E(XY) = E(X) \times E(Y),$$ thus:

$$\mathrm{cov}(X,Y) = E(XY) - \mu_x\mu_y = E(X) \times E(Y) - \mu_x\mu_y = \mu_x\mu_y - \mu_x\mu_y = 0$$

The two variables are thus also not correlated, which is reasonable.

Interestingly, while two independent variables are certainly not correlated, two non correlated variables are not necessarily independent.

For example, given variables X, uniformly distributed on values {-1, 0, 1}, and Y =|X|, their product is:

$XY = X|X| = X$, thus $E(XY) = E(X) = 0$,

whereas $E(Y)= (1/3) * |-1| + (1/3) * 0 + (1/3) * |1| = 2/3$.

In this case, $\mathrm{cov}(X,Y) = E(XY) - E(X) \times E(Y) = 0$, thus X and Y are not correlated.

However, X and Y are not independent.

Indeed:

$P(X=1, Y = 0) = 0$, by definition of Y

$P(X=1) \times P(Y =0) = 1/3 \times 1/3 = 1/9$

Consequently:

$P(X=1, Y = 0) \neq P(X=1) \times P(Y =0)$

Covariance has to be considered in the general calculation of the variance related to the sum of two random variables:

$$V[X+Y] = E\{[(x+y) - (\mu_x + \mu_y)]^2\} = E\{[(x-\mu_x) + (y-\mu_y)]^2\} =$$

$$= E\{(x-\mu_x)^2\} + E\{(y-\mu_y)^2\} + 2E\{(x-\mu_x)(y-\mu_y)\} =$$

$$= V(X) + V(Y) + 2\,Cov\,(X,Y)$$

By analogy:

$$V[X-Y] = V(X) + V(Y) - 2\,Cov\,(X,Y)$$

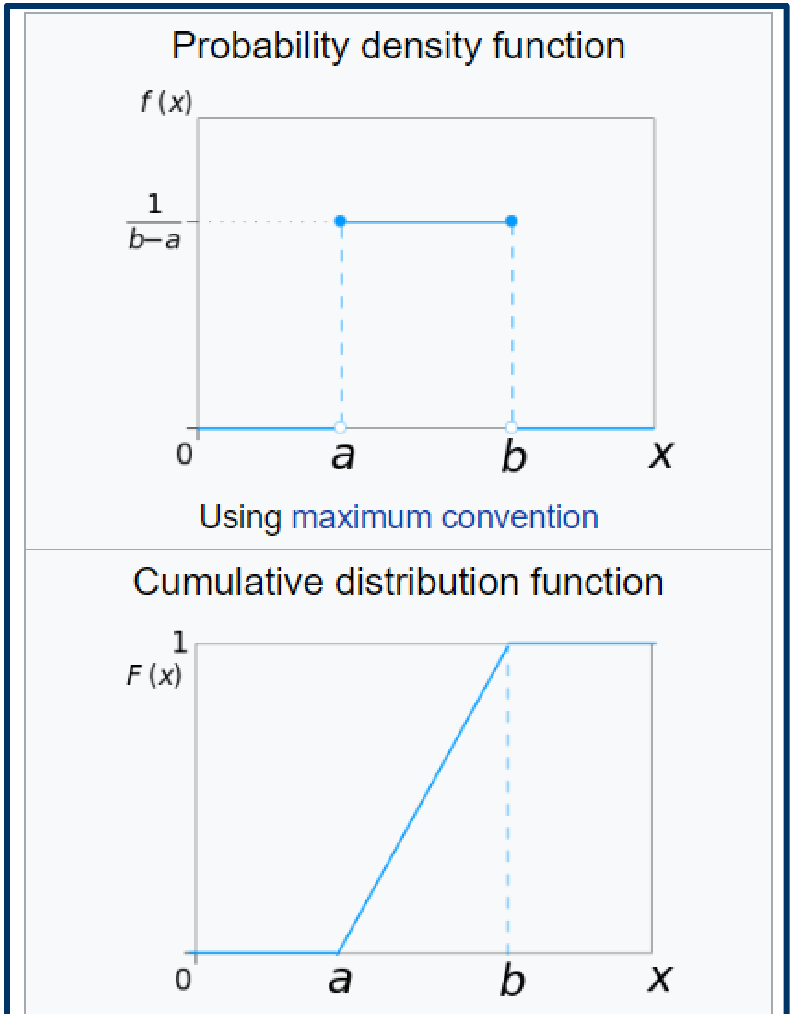
These relations can be extended easily to the case of variance related to the sum of n random variables:

$$V[\textstyle\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} V(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov[Xi, Yi]$$

If all random variables are independent (and, then, also not correlated) the variance of their sum simply corresponds to the sum of their variances.

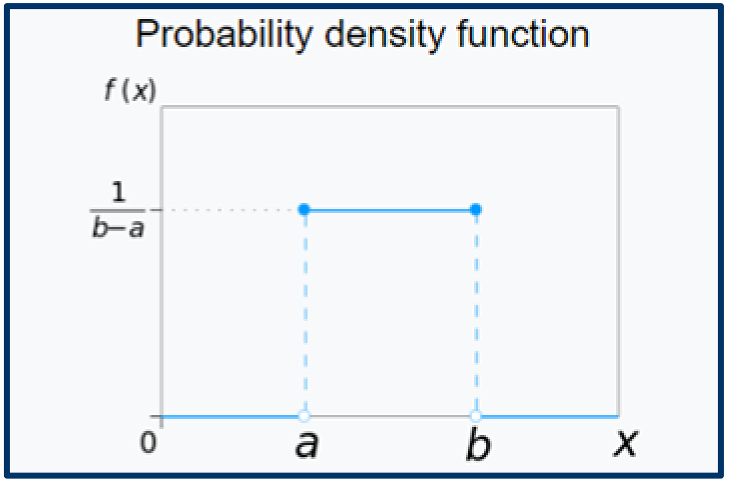
Examples of distributions

Uniform distribution (continuous)



Notation	$\mathcal{U}(a,b)$ or $\mathrm{unif}(a,b)$
Parameters	$0 < a < b < \infty$
Support	$x \in [a,b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b) \\ 1 & \text{for } x \geq b \end{cases}$
Mean	$\frac{1}{2}(a+b)$
Median	$\frac{1}{2}(a+b)$
Mode	any value in (a,b)
Variance	$\frac{1}{12}(b-a)^2$
Skewness	0
Ex. kurtosis	$-\frac{6}{5}$
Entropy	$\ln(b-a)$
MGF	$\begin{cases} \frac{e^{tb}-e^{ta}}{t(b-a)} & \text{for } t \neq 0 \\ 1 & \text{for } t = 0 \end{cases}$

It can be easily demonstrated that the value assumed by f(x) for x comprised between a and b is 1/(b-a).



Indeed, if k corresponds to that value, the following equations can be written:

$$\int_{0}^{+\infty} f(x)dx = \int_{a}^{b} kdx = k\int_{a}^{b} dx = k\,(b-a) = 1$$

Hence: k = 1/(b-a).

The population mean can be calculated as follows:

$$\mu = E(x) = \int_0^{+\infty} x\, f(x)dx = \int_a^b xkdx = k\int_a^b xdx = \frac{k}{2}\,(b^2-a^2) = \frac{1}{2}\frac{(b+a)(b-a)}{(b-a)} =$$

$$= \frac{1}{2}(b+a)$$

The population variance can be calculated as follows:

$$V(x) = E\{(x-\mu)^2\} = E\{x^2\} - \{E(x)\}^2$$

Since:

$$E\{x^2\} = \int_0^{+\infty} x^2 f(x)dx = \int_a^b kx^2 dx = \frac{k}{3}(b^3 - a^3) = \frac{1}{3}\frac{(b-a)(b^2 + a^2 + ab)}{(b-a)}$$

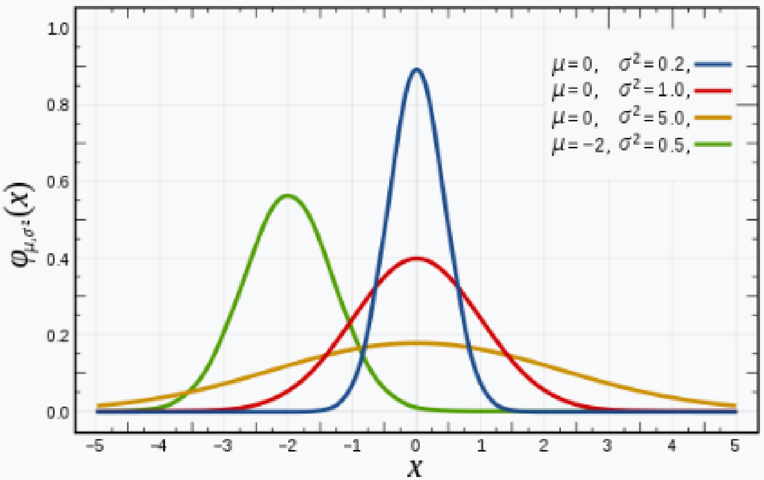
and:

$$\{E(x)\}^2 = \frac{(b+a)^2}{4} = \frac{b^2 + a^2 + 2\,ab}{4}$$

$$V(x) = \frac{(b^2 + a^2 + ab)}{3} - \frac{b^2 + a^2 + 2\,ab}{4} = \frac{b^2 + a^2 - 2\,ab}{12} = \frac{(b-a)^2}{12}$$

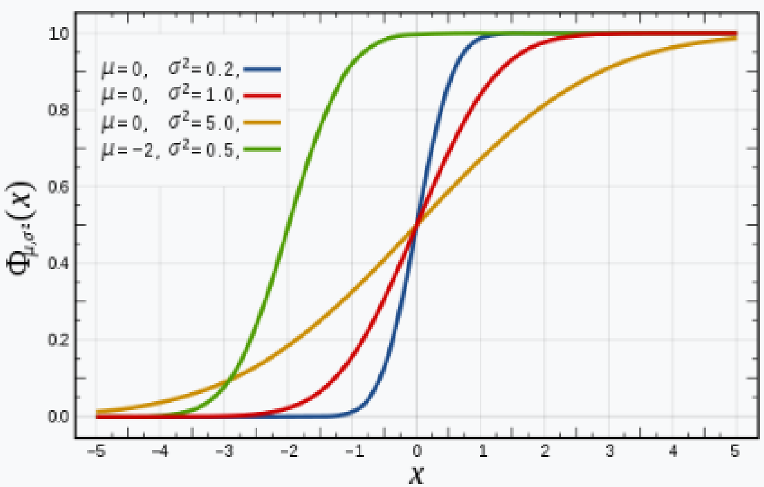
Normal (Gaussian) distribution

Probability density function



The red curve is the *standard normal distribution*

Cumulative distribution function



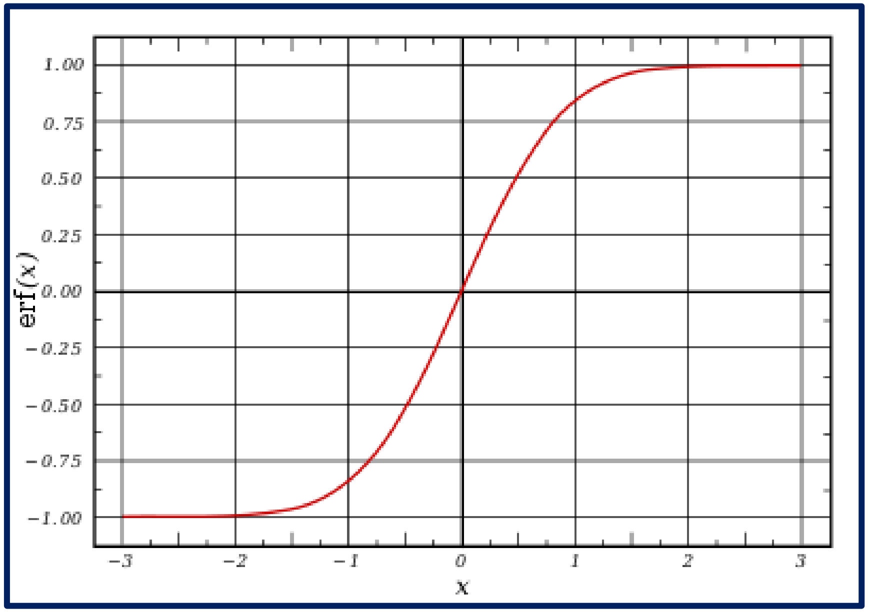
Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\operatorname{erf}^{-1}(2F-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
Skewness	0
Ex. kurtosis	0
Entropy	$\frac{1}{2}\log(2\pi e\sigma^2)$
MGF	$\exp(\mu t + \sigma^2 t^2/2)$

Relationship between Normal Cumulative Distribution Function (CDF) and Error function (erf)

By definition, the error function of x is:

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t$$

which corresponds to a sigmoidal function:



In order to find the relationship between Normal CDF and erf a variable replacement has to be made:

$$t = z/\sqrt{2} \Rightarrow \mathrm{d}t = \mathrm{d}z/\sqrt{2}$$

Consequently:

$$t = 0 \Rightarrow z = 0 \quad \text{and} \quad t = x \Rightarrow z = x\sqrt{2}$$

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{2\pi}} \int_0^{x\sqrt{2}} e^{-z^2/2} \mathrm{d}z = 2\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x\sqrt{2}} e^{-z^2/2} \mathrm{d}z - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-z^2/2} \mathrm{d}z\right)$$

The two integrals reported in the right side of the equation correspond to specific values of the CDF for the standard normal distribution (i.e., of the function $e^{-z^2/2}$):

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

Consequently:

$$\mathrm{Erf}(x) = 2(\Phi(x\sqrt{2}) - \Phi(0)) = 2\left(\Phi(x\sqrt{2}) - \frac{1}{2}\right) = 2\Phi(x\sqrt{2}) - 1$$

then:

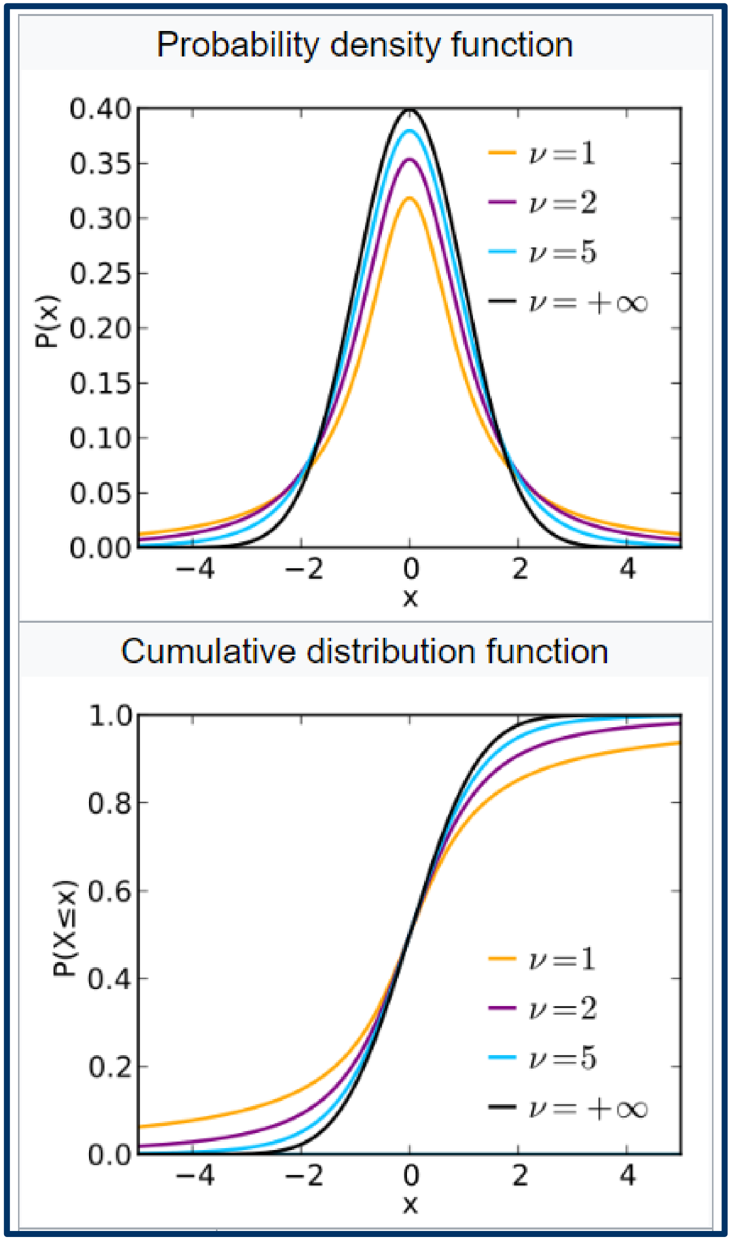
$$\Phi(x) = \frac{1 + \mathrm{Erf}(x/\sqrt{2})}{2}$$

If the CDF of a general (not of the standard) normal distribution is considered, variable x in the above equation (equivalent to z) needs to be replaced by (x-μ)/σ, thus the following general equation is obtained:

$$\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

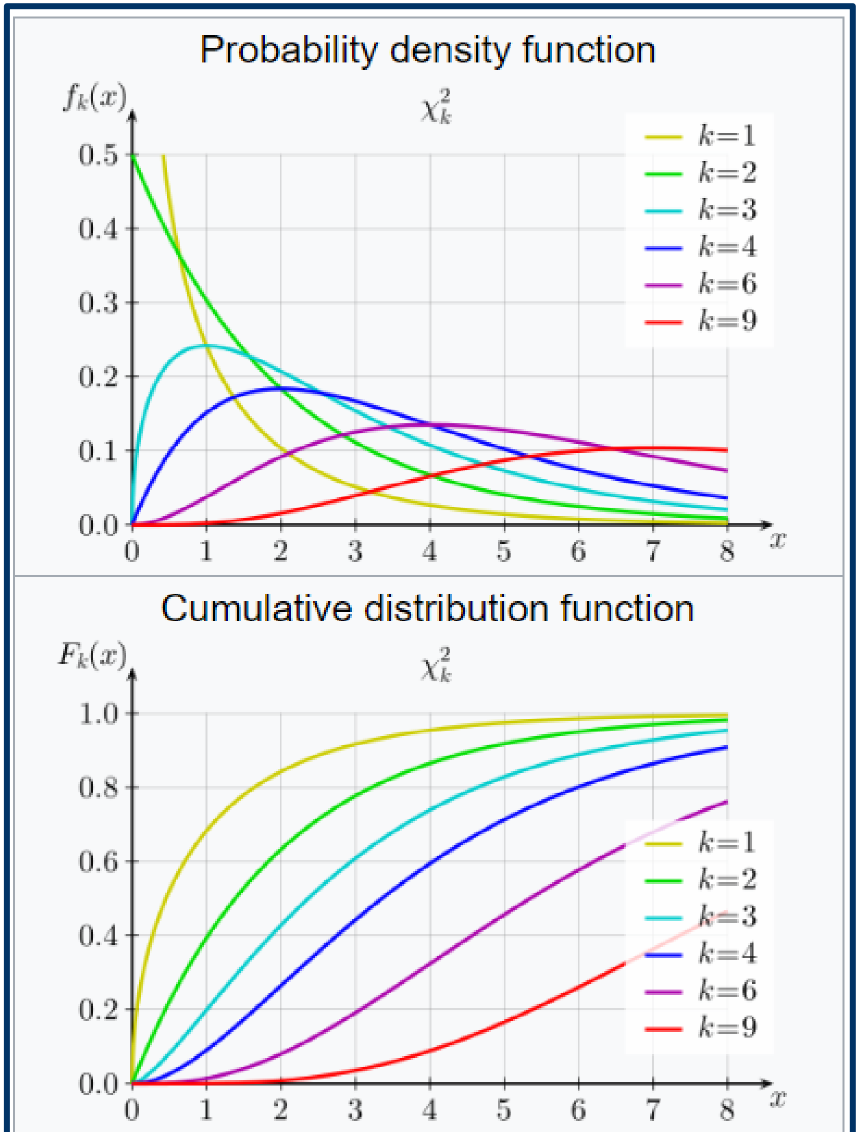
Since an erf is a sigmoidal function and Φ(x) is expressed as the sum between an erf and a constant, it is not surprising that also Φ(x) has a sigmoidal shape.

Student's t distribution



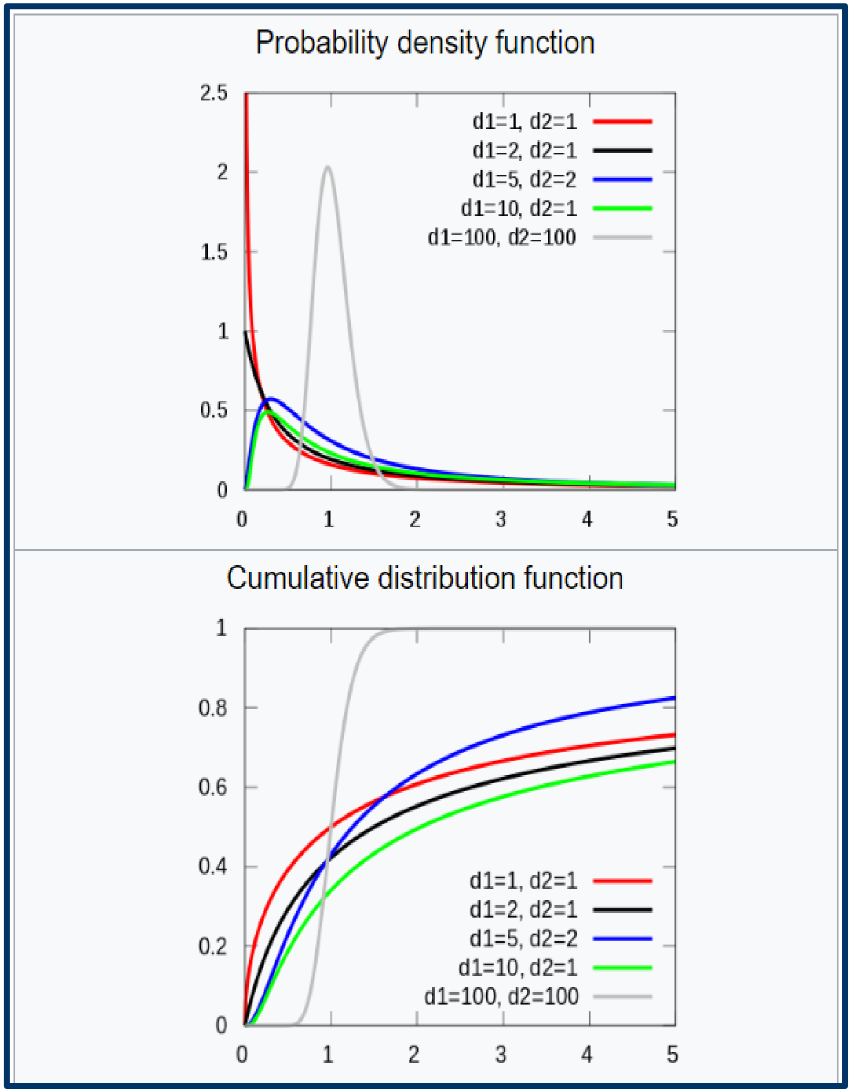
Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty, \infty)$
PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$ where ${}_2F_1$ is the hypergeometric function
Mean	0 for $\nu > 1$, otherwise undefined
Median	0
Mode	0
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$, ∞ for $1 < \nu \le 2$, otherwise undefined
Skewness	0 for $\nu > 3$, otherwise undefined
Ex. kurtosis	$\frac{6}{\nu-4}$ for $\nu > 4$, ∞ for $2 < \nu \le 4$, otherwise undefined

χ^2 (chi-squared) distribution



Notation	$\chi^2(k)$ or χ_k^2
Parameters	$k \in \mathbb{N}_{>0}$ (known as "degrees of freedom")
Support	$x \in (0, +\infty)$ if $k = 1$, otherwise $x \in [0, +\infty)$
PDF	$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(k/2)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$
Mean	k
Median	$\approx k\left(1 - \frac{2}{9k}\right)^3$
Mode	$\max(k-2, 0)$
Variance	$2k$
Skewness	$\sqrt{8/k}$
Ex. kurtosis	$\frac{12}{k}$
Entropy	$\frac{k}{2} + \ln(2\Gamma(\frac{k}{2})) + (1-\frac{k}{2})\psi(\frac{k}{2})$

Fisher-Snedecor (F) distribution



Parameters	d_1, d_2 > 0 deg. of freedom
Support	$x \in (0, +\infty)$ if $d_1 = 1$, otherwise $x \in [0, +\infty)$
PDF	$\dfrac{\sqrt{\dfrac{(d_1\,x)^{d_1}\,d_2^{d_2}}{(d_1\,x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$
CDF	$I_{\frac{d_1 x}{d_1 x+d_2}}\left(\frac{d_1}{2},\frac{d_2}{2}\right)$
Mean	$\dfrac{d_2}{d_2-2}$ for d_2 > 2
Mode	$\dfrac{d_1-2}{d_1}\,\dfrac{d_2}{d_2+2}$ for d_1 > 2
Variance	$\dfrac{2\,d_2^2\,(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ for d_2 > 4
Skewness	$\dfrac{(2d_1+d_2-2)\sqrt{8(d_2-4)}}{(d_2-6)\sqrt{d_1(d_1+d_2-2)}}$ for d_2 > 6