Beta distribution

From Wikipedia, the free encyclopedia

In probability theory and statistics, the **beta distribution** is a family of continuous probability distributions defined on the interval [0, 1] parametrized by two positive shape parameters, denoted by α and β, that appear as exponents of the random variable and control the shape of the distribution.

The beta distribution has been applied to model the behavior of random variables limited to intervals of finite length in a wide variety of disciplines. For example, it has been used as a statistical description of allele frequencies in population genetics;[1] time allocation in project management / control systems;[2] sunshine data;[3] variability of soil properties;[4] proportions of the minerals in rocks in stratigraphy;[5] and heterogeneity in the probability of HIV transmission.[6]

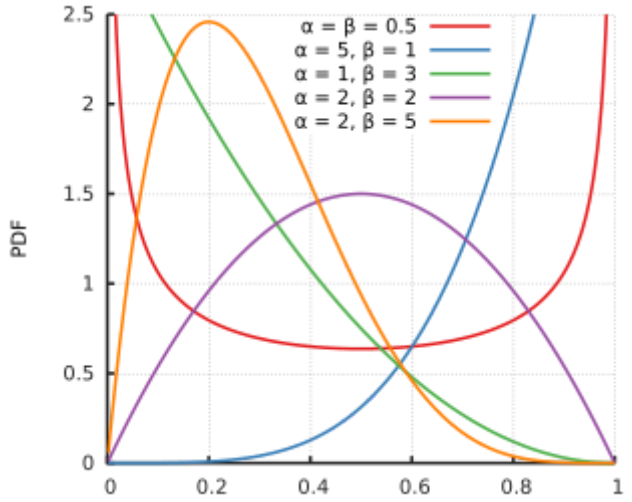
In Bayesian inference, the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions. For example, the beta distribution can be used in Bayesian analysis to describe initial knowledge concerning probability of success such as the probability that a space vehicle will successfully complete a specified mission. The beta distribution is a suitable model for the random behavior of percentages and proportions.

The usual formulation of the beta distribution is also known as the **beta distribution of the first kind**, whereas *beta distribution of the second kind* is an alternative name for the beta prime distribution.

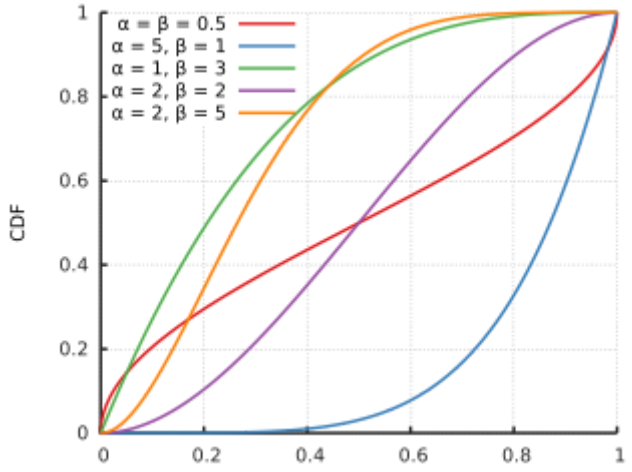
Contents

Beta

Probability density function



Cumulative distribution function



Notation	Beta(α, β)

Parameters	$\alpha > 0$ shape (real) $\beta > 0$ shape (real)
Support	$x \in [0,1]$ or $x \in (0,1)$
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$ where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
CDF	$I_x(\alpha,\beta)$
Mean	$\mathrm{E}[X] = \frac{\alpha}{\alpha+\beta}$ $\mathrm{E}[\ln X] = \psi(\alpha) - \psi(\alpha+\beta)$ (see digamma function and see section: Geometric mean)
Median	$I^{[-1]}_{\frac{1}{2}}(\alpha,\beta)$ **(in general)** $\approx \frac{\alpha-\frac{1}{3}}{\alpha+\beta-\frac{2}{3}}$ **for** $\alpha,\beta > 1$
Mode	$\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$
Variance	$\mathrm{var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ $\mathrm{var}[\ln X] = \psi_1(\alpha) - \psi_1(\alpha+\beta)$ (see trigamma function and see section: Geometric variance)
Skewness	$\frac{2\,(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$
Ex. kurtosis	$\frac{6[(\alpha-\beta)^2(\alpha+\beta+1) - \alpha\beta(\alpha+\beta+2)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$
Entropy	$\ln \mathrm{B}(\alpha,\beta) - (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta)$ $+(\alpha+\beta-2)\psi(\alpha+\beta)$

MGF	$1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$
CF	${}_1F_1(\alpha;\alpha+\beta;i\,t)$ (see Confluent hypergeometric function)
Fisher information	$\begin{matrix}\operatorname{var}[\ln X] & \operatorname{cov}[\ln X,\ln(1-X)]\\ \operatorname{cov}[\ln X,\ln(1-X)] & \operatorname{var}[\ln(1-X)]\end{matrix}$ see section: Fisher information matrix

Characterization

Probability density function

The probability density function (pdf) of the beta distribution, for $0 \le x \le 1$, and shape parameters $\alpha, \beta > 0$, is a power function of the variable x and of its reflection $(1 - x)$ as follows:

$$\begin{aligned} f(x;\alpha,\beta) &= \text{constant} \cdot x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \end{aligned}$$

where $\Gamma(z)$ is the gamma function. The beta function, B, is a normalization constant to ensure that the total probability integrates to 1. In the above equations x is a realization—an observed value that actually occurred—of a random process X.

This definition includes both ends $x = 0$ and $x = 1$, which is consistent with definitions for other continuous distributions supported on a bounded interval which are special cases of the beta distribution, for example the arcsine distribution, and consistent with several authors, like N. L. Johnson and S. Kotz.[7][8][9][10] However, the inclusion of $x = 0$ and $x = 1$ does not work for $\alpha, \beta < 1$; accordingly, several other authors, including W. Feller,[11][12][13] choose to exclude the ends $x = 0$ and $x = 1$, (such that the two ends are not actually part of the density function) and consider instead $0 < x < 1$.

Several authors, including N. L. Johnson and S. Kotz,[7] use the symbols p and q (instead of α and β) for the shape parameters of the beta distribution, reminiscent of the symbols traditionally used for the parameters of the Bernoulli distribution, because the beta distribution approaches the Bernoulli distribution in the limit when both shape parameters α and β approach the value of zero.

In the following, a random variable X beta-distributed with parameters α and β will be denoted by:[14][15]

$$X \sim \mathrm{Beta}(\alpha,\beta)$$

Other notations for beta-distributed random variables used in the statistical literature are $X \sim \mathcal{B}e(\alpha,\beta)$[16] and $X \sim \beta_{\alpha,\beta}$.[11]

Differential equation

The probability density function satisfies the differential equation

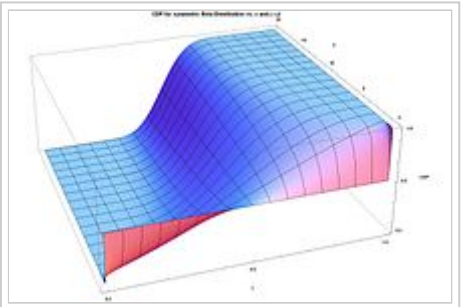
$$f'(x) = f(x)\,\frac{(\alpha+\beta-2)x-(\alpha-1)}{(x-1)x}.$$

Cumulative distribution function

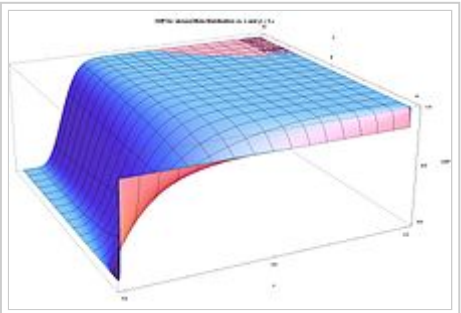
The cumulative distribution function is

$$F(x;\alpha,\beta) = \frac{\mathrm{B}(x;\alpha,\beta)}{\mathrm{B}(\alpha,\beta)} = I_x(\alpha,\beta)$$

where $\mathrm{B}(x;\alpha,\beta)$ is the incomplete beta function and $I_x(\alpha,\beta)$ is the regularized incomplete beta function.



CDF for symmetric beta distribution vs. x and alpha=beta



CDF for skewed beta distribution vs. x and beta= 5 alpha

Properties

Measures of central tendency

Mode

The mode of a Beta distributed random variable X with $\alpha, \beta > 1$ is the most likely value of the distribution (corresponding to the peak in the PDF), and is given by the following expression:[7]

$$\frac{\alpha-1}{\alpha+\beta-2}.$$

When both parameters are less than one ($\alpha, \beta < 1$), this is the anti-mode: the lowest point of the probability density curve.[9]

Letting $\alpha = \beta$, the expression for the mode simplifies to 1/2, showing that for $\alpha = \beta > 1$ the mode (resp. anti-mode when $\alpha, \beta < 1$), is at the center of the distribution: it is symmetric in those cases. See "Shapes" section in this article for a full list of mode cases, for arbitrary values of α and β. For several of these cases, the maximum value of the density function occurs at one or both ends. In some cases the (maximum) value of the density function occurring at the end is finite. For example, in the case of $\alpha = 2, \beta = 1$ (or $\alpha = 1, \beta = 2$), the density function becomes a right-triangle distribution which is finite at both ends. In several other cases there is a singularity at one end, where the value of the density function approaches infinity. For example, in the case $\alpha = \beta = 1/2$, the Beta distribution simplifies to become the arcsine distribution. There is debate among mathematicians about some of these cases and whether the ends ($x = 0$, and $x = 1$) can be called *modes* or not.[12][14]

- Whether the ends are part of the domain of the density function
- Whether a singularity can ever be called a *mode*
- Whether cases with two maxima should be called *bimodal*

Median

The median of the beta distribution is the unique real number $x = I^{[-1]}_{\frac{1}{2}}(\alpha, \beta)$ for which the regularized incomplete beta function $I_x(\alpha, \beta) = \frac{1}{2}$. There is no general closed-form expression for the median of the beta distribution for arbitrary values of α and β. Closed-form expressions for particular values of the parameters α and β follow:

- For symmetric cases α = β, median = 1/2.
- For α = 1 and β > 0, median = $\mathbf{1 - 2^{-\frac{1}{\beta}}}$ (this case is the mirror-image of the power function [0,1] distribution)
- For α > 0 and β = 1, median = $\mathbf{2^{-\frac{1}{\alpha}}}$ (this case is the power function [0,1] distribution[12])
- For α = 3 and β = 2, median = 0.6142724318676105..., the real solution to the quartic equation $1-8x^3+6x^4 = 0$, which lies in [0,1].
- For α = 2 and β = 3, median = 0.38572756813238945... = 1−median(Beta(3, 2))

The following are the limits with one parameter finite (non zero) and the other approaching these limits:

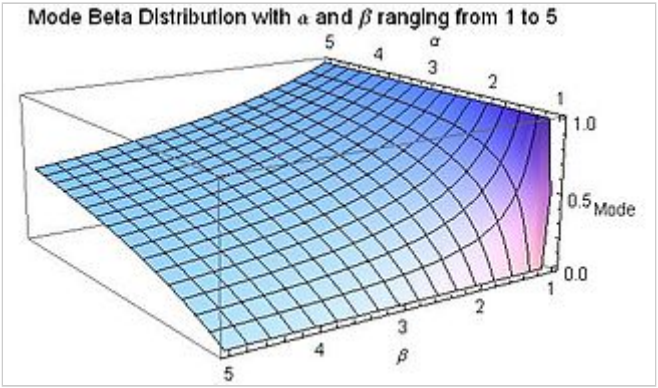
$$\lim_{\beta \to 0} \mathbf{median} = \lim_{\alpha \to \infty} \mathbf{median} = \mathbf{1},$$

$$\lim_{\alpha \to 0} \mathbf{median} = \lim_{\beta \to \infty} \mathbf{median} = \mathbf{0}.$$

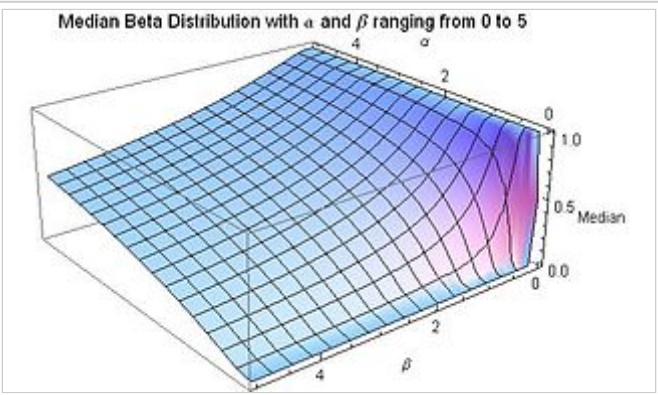
A reasonable approximation of the value of the median of the beta distribution, for both α and β greater or equal to one, is given by the formula[17]

$$\mathbf{median} \approx \frac{\alpha - \frac{1}{3}}{\alpha + \beta - \frac{2}{3}} \mathbf{\ for\ } \alpha, \beta \geq 1.$$

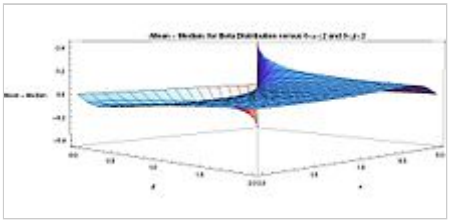
When α, β ≥ 1, the relative error (the absolute error divided by the median) in this approximation is less than 4% and for both α ≥ 2 and β ≥ 2 it is less than 1%. The absolute error divided by the difference between the mean and the mode is similarly small:

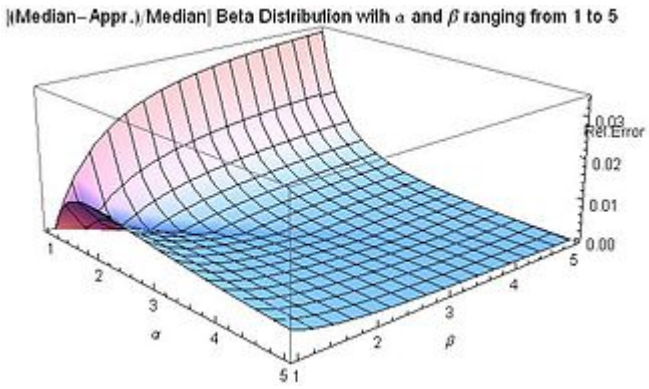


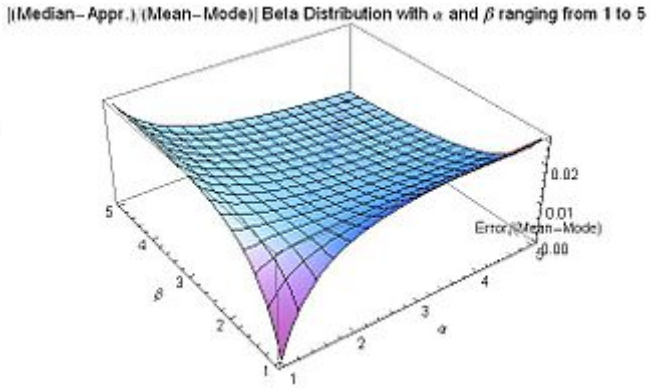
Mode for Beta distribution for 1 ≤ α ≤ 5 and 1 ≤ β ≤ 5



Median for Beta distribution for 0≤α≤5 and 0≤β≤5


(Mean - Median) for Beta distribution versus alpha and beta from 0 to 2

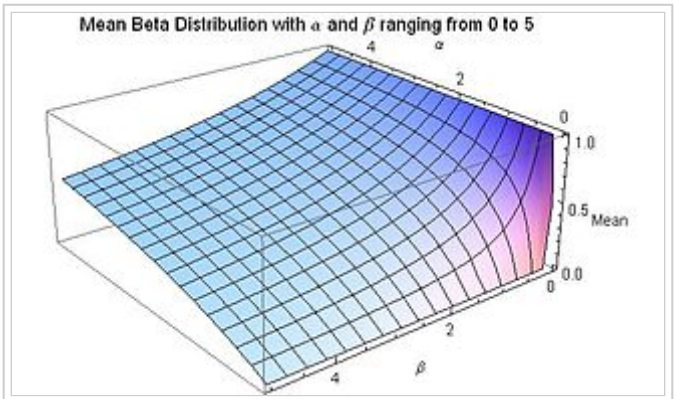




Mean

The expected value (mean) (μ) of a Beta distribution random variable X with two parameters α and β is a function of only the ratio β/α of these parameters:[7]

$$\begin{aligned}\mu = \mathrm{E}[X] &= \int_0^1 x f(x;\alpha,\beta)\,dx \\ &= \int_0^1 x\,\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,dx \\ &= \frac{\alpha}{\alpha+\beta} \\ &= \frac{1}{1+\frac{\beta}{\alpha}}\end{aligned}$$



Mean for Beta distribution for $0 \le \alpha \le 5$ and $0 \le \beta \le 5$

Letting $\alpha = \beta$ in the above expression one obtains $\mu = 1/2$, showing that for $\alpha = \beta$ the mean is at the center of the distribution: it is symmetric. Also, the following limits can be obtained from the above expression:

$$\lim_{\frac{\beta}{\alpha}\to 0} \mu = 1$$

$$\lim_{\frac{\beta}{\alpha}\to \infty} \mu = 0$$

Therefore, for $\beta/\alpha \to 0$, or for $\alpha/\beta \to \infty$, the mean is located at the right end, $x = 1$. For these limit ratios, the beta distribution becomes a one-point degenerate distribution with a Dirac delta function spike at the right end, $x = 1$, with probability 1, and zero probability everywhere else. There is 100% probability (absolute certainty) concentrated at the right end, $x = 1$.

Similarly, for $\beta/\alpha \to \infty$, or for $\alpha/\beta \to 0$, the mean is located at the left end, $x = 0$. The beta distribution becomes a 1-point Degenerate distribution with a Dirac delta function spike at the left end, $x = 0$, with probability 1, and zero probability everywhere else. There is 100% probability (absolute certainty) concentrated at the left end, $x = 0$. Following are the limits with one parameter finite (non zero) and the other approaching these limits:

$$\lim_{\beta\to 0} \mu = \lim_{\alpha\to\infty} \mu = 1$$
$$\lim_{\alpha\to 0} \mu = \lim_{\beta\to\infty} \mu = 0$$

While for typical unimodal distributions (with centrally located modes, inflexion points at both sides of the mode, and longer tails) (with Beta(α, β) such that $\alpha, \beta > 2$) it is known that the sample mean (as an estimate of location) is not as robust as the sample median, the opposite is the case for uniform or "U-shaped" bimodal distributions (with Beta(α, β) such that $\alpha, \beta \leq 1$), with the modes located at the ends of the distribution. As Mosteller and Tukey remark ([18] p. 207) "the average of the two extreme observations uses all the sample information. This illustrates how, for short-tailed distributions, the extreme observations should get more weight." By contrast, it follows that the median of "U-shaped" bimodal distributions with modes at the edge of the distribution (with Beta(α, β) such that $\alpha, \beta \leq 1$) is not robust, as the sample median drops the extreme sample observations from consideration. A practical application of this occurs for example for random walks, since the probability for the time of the last visit to the origin in a random walk is distributed as the arcsine distribution Beta(1/2, 1/2):[11][19] the mean of a number of realizations of a random walk is a much more robust estimator than the median (which is an inappropriate sample measure estimate in this case).

Geometric mean

The logarithm of the geometric mean G_X of a distribution with random variable X is the arithmetic mean of ln(X), or, equivalently, its expected value:

$$\ln G_X = \mathbf{E}[\ln X]$$

For a beta distribution, the expected value integral gives:

$$\begin{aligned}
\mathrm{E}[\ln X] &= \int_0^1 \ln x\, f(x;\alpha,\beta)\, dx \\
&= \int_0^1 \ln x\, \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\, dx \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 \frac{\partial x^{\alpha-1}(1-x)^{\beta-1}}{\partial \alpha}\, dx \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)} \frac{\partial}{\partial \alpha} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)} \frac{\partial \mathrm{B}(\alpha,\beta)}{\partial \alpha} \\
&= \frac{\partial \ln \mathrm{B}(\alpha,\beta)}{\partial \alpha} \\
&= \frac{\partial \ln \Gamma(\alpha)}{\partial \alpha} - \frac{\partial \ln \Gamma(\alpha+\beta)}{\partial \alpha} \\
&= \psi(\alpha) - \psi(\alpha+\beta)
\end{aligned}$$

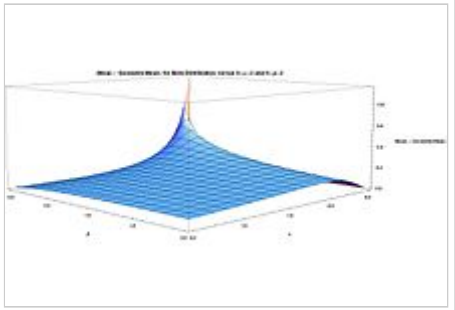
where ψ is the digamma function.

Therefore, the geometric mean of a beta distribution with shape parameters α and β is the exponential of the digamma functions of α and β as follows:

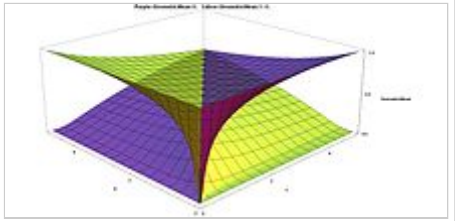
$$G_X = e^{\mathrm{E}[\ln X]} = e^{\psi(\alpha)-\psi(\alpha+\beta)}$$

While for a beta distribution with equal shape parameters α = β, it follows that skewness = 0 and mode = mean = median = 1/2, the geometric mean is less than 1/2: $0 < G_X < 1/2$. The reason for this is that the logarithmic transformation strongly weights the values of X close to zero, as $\ln(X)$ strongly tends towards negative infinity as X approaches zero, while $\ln(X)$ flattens towards zero as $X \to 1$.

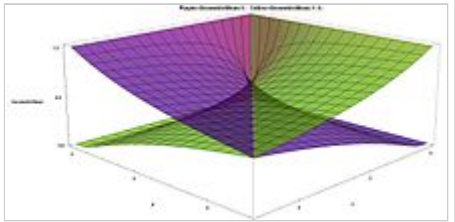
Along a line $\alpha = \beta$, the following limits apply:



(Mean − GeometricMean) for Beta distribution versus α and β from 0 to 2, showing the asymmetry between α and β for the geometric mean



Geometric means for Beta distribution Purple = $G(x)$, Yellow = $G(1-x)$, smaller values alpha and beta in front



Geometric means for Beta distribution. purple = $G(x)$, yellow = $G(1-x)$, larger values α and β in front

$$\lim_{\alpha=\beta\to 0} G_X = 0$$
$$\lim_{\alpha=\beta\to \infty} G_X = \tfrac{1}{2}$$

Following are the limits with one parameter finite (non zero) and the other approaching these limits:

$$\lim_{\beta\to 0} G_X = \lim_{\alpha\to\infty} G_X = 1$$
$$\lim_{\alpha\to 0} G_X = \lim_{\beta\to\infty} G_X = 0$$

The accompanying plot shows the difference between the mean and the geometric mean for shape parameters α and β from zero to 2. Besides the fact that the difference between them approaches zero as α and β approach infinity and that the difference becomes large for values of α and β approaching zero, one can observe an evident asymmetry of the geometric mean with respect to the shape parameters α and β. The difference between the geometric mean and the mean is larger for small values of α in relation to β than when exchanging the magnitudes of β and α.

N. L.Johnson and S. Kotz[7] suggest the logarithmic approximation to the digamma function $\psi(\alpha) \approx \ln(\alpha-1/2)$ which results in the following approximation to the geometric mean:

$$G_X \approx \frac{\alpha - \frac{1}{2}}{\alpha + \beta - \frac{1}{2}} \text{ if } \alpha, \beta > 1.$$

Numerical values for the relative error in this approximation follow: [$(\alpha = \beta = 1)$: 9.39%]; [$(\alpha = \beta = 2)$: 1.29%]; [$(\alpha = 2, \beta = 3)$: 1.51%]; [$(\alpha = 3, \beta = 2)$: 0.44%]; [$(\alpha = \beta = 3)$: 0.51%]; [$(\alpha = \beta = 4)$: 0.26%]; [$(\alpha = 3, \beta = 4)$: 0.55%]; [$(\alpha = 4, \beta = 3)$: 0.24%].

Similarly, one can calculate the value of shape parameters required for the geometric mean to equal 1/2. Given the value of the parameter β, what would be the value of the other parameter, α, required for the geometric mean to equal 1/2?. The answer is that (for $\beta > 1$), the value of α required tends towards $\beta + 1/2$ as $\beta \to \infty$. For example, all these couples have the same geometric mean of 1/2: [$\beta = 1$, $\alpha = 1.4427$], [$\beta = 2$, $\alpha = 2.46958$], [$\beta = 3$, $\alpha = 3.47943$], [$\beta = 4$, $\alpha = 4.48449$], [$\beta = 5$, $\alpha = 5.48756$], [$\beta = 10$, $\alpha = 10.4938$], [$\beta = 100$, $\alpha = 100.499$].

The fundamental property of the geometric mean, which can be proven to be false for any other mean, is

$$G\left(\frac{X_i}{Y_i}\right) = \frac{G(X_i)}{G(Y_i)}$$

This makes the geometric mean the only correct mean when averaging *normalized* results, that is results that are presented as ratios to reference values.[20] This is relevant because the beta distribution is a suitable model for the random behavior of percentages and it is particularly suitable to the statistical modelling of proportions. The geometric mean plays a central role in maximum likelihood estimation, see section "Parameter estimation, maximum likelihood." Actually, when performing maximum likelihood estimation, besides the geometric mean G_X based on the random variable X, also another geometric mean appears naturally: the geometric mean based on the linear transformation —(1 − X), the mirror-image of X, denoted by $G_{(1-X)}$:

$$G_{(1-X)} = e^{\mathrm{E}[\ln(1-X)]} = e^{\psi(\beta)-\psi(\alpha+\beta)}$$

Along a line $\alpha = \beta$, the following limits apply:

$$\lim_{\alpha=\beta\to 0} G_{(1-X)} = 0$$
$$\lim_{\alpha=\beta\to \infty} G_{(1-X)} = \tfrac{1}{2}$$

Following are the limits with one parameter finite (non-zero) and the other approaching these limits:

$$\lim_{\beta\to 0} G_{(1-X)} = \lim_{\alpha\to\infty} G_{(1-X)} = 0$$
$$\lim_{\alpha\to 0} G_{(1-X)} = \lim_{\beta\to\infty} G_{(1-X)} = 1$$

It has the following approximate value:

$$G_{(1-X)} \approx \frac{\beta - \frac{1}{2}}{\alpha+\beta-\frac{1}{2}} \text{ if } \alpha, \beta > 1.$$

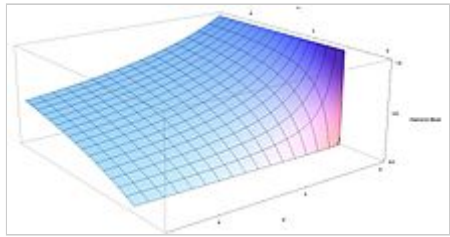
Although both G_X and $G_{(1-X)}$ are asymmetric, in the case that both shape parameters are equal $\alpha = \beta$, the geometric means are equal: $G_X = G_{(1-X)}$. This equality follows from the following symmetry displayed between both geometric means:

$$G_X(\mathrm{B}(\alpha,\beta)) = G_{(1-X)}(\mathrm{B}(\beta,\alpha)).$$

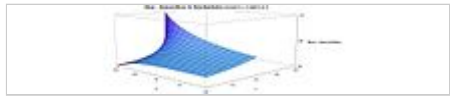
Harmonic mean

The inverse of the harmonic mean (H_X) of a distribution with random variable X is the arithmetic mean of 1/X, or, equivalently, its expected value. Therefore, the harmonic mean (H_X) of a beta distribution with shape parameters α and β is:

$$H_X = \frac{1}{\mathrm{E}\left[\frac{1}{X}\right]} = \frac{1}{\int_0^1 \frac{f(x;\alpha,\beta)}{x}\,dx} = \frac{1}{\int_0^1 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{x\mathrm{B}(\alpha,\beta)}\,dx} = \frac{\alpha-1}{\alpha+\beta-1} \text{ if } \alpha > 1 \text{ and } \beta > 0$$



Harmonic mean for Beta distribution for 0<α<5 and 0<β<5



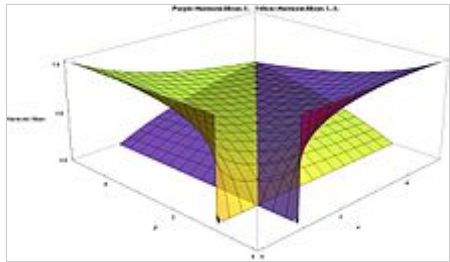
(Mean - HarmonicMean) for Beta distribution versus alpha and beta from 0 to 2

The harmonic mean (H_X) of a Beta distribution with α < 1 is undefined, because its defining expression is not bounded in [0, 1] for shape parameter α less than unity.

Letting α = β in the above expression one obtains

$$H_X = \frac{\alpha-1}{2\alpha-1},$$

showing that for α = β the harmonic mean ranges from 0, for α = β = 1, to 1/2, for α = β → ∞.



Harmonic Means for Beta distribution Purple=H(X), Yellow=H(1-X), smaller values alpha and beta in front

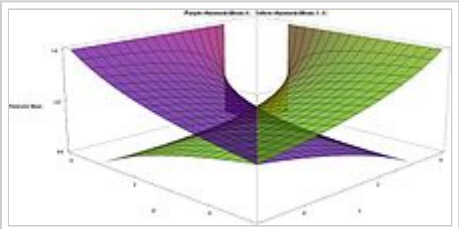
Following are the limits with one parameter finite (non zero) and the other approaching these limits:

$$\lim_{\alpha\to 0} H_X = \text{undefined}$$
$$\lim_{\alpha\to 1} H_X = \lim_{\beta\to\infty} H_X = 0$$
$$\lim_{\beta\to 0} H_X = \lim_{\alpha\to\infty} H_X = 1$$

The harmonic mean plays a role in maximum likelihood estimation for the four parameter case, in addition to the geometric mean. Actually, when performing maximum likelihood estimation for the four parameter case, besides the harmonic mean H_X based on the random variable X, also another harmonic mean appears naturally: the harmonic mean based on the linear transformation $(1-X)$, the mirror-image of X, denoted by $H_{(1-X)}$:

$$H_{(1-X)} = \frac{1}{\mathrm{E}[\frac{1}{(1-X)}]} = \frac{\beta-1}{\alpha+\beta-1} \text{ if } \beta > 1, \&\alpha > 0.$$

The harmonic mean ($H_{(1-X)}$) of a Beta distribution with β < 1 is undefined, because its defining expression is not bounded in [0, 1] for shape parameter β less than unity.



Harmonic Means for Beta distribution Purple=H(X), Yellow=H(1-X), larger values alpha and beta in front

Letting α = β in the above expression one obtains

$$H_{(1-X)} = \frac{\beta - 1}{2\beta - 1},$$

showing that for α = β the harmonic mean ranges from 0, for α = β = 1, to 1/2, for α = β → ∞.

Following are the limits with one parameter finite (non zero) and the other approaching these limits:

$$\lim_{\beta \to 0} H_{(1-X)} = \text{undefined}$$
$$\lim_{\beta \to 1} H_{(1-X)} = \lim_{\alpha \to \infty} H_{(1-X)} = 0$$
$$\lim_{\alpha \to 0} H_{(1-X)} = \lim_{\beta \to \infty} H_{(1-X)} = 1$$

Although both H_X and $H_{(1-X)}$ are asymmetric, in the case that both shape parameters are equal α = β, the harmonic means are equal: $H_X = H_{(1-X)}$. This equality follows from the following symmetry displayed between both harmonic means:

$$H_X(\mathrm{B}(\alpha, \beta)) = H_{(1-X)}(\mathrm{B}(\beta, \alpha)) \text{ if } \alpha, \beta > 1.$$

Measures of statistical dispersion

Variance

The variance (the second moment centered on the mean) of a Beta distribution random variable X with parameters α and β is:[7][21]

$$\operatorname{var}(X) = \mathrm{E}[(X - \mu)^2] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Letting α = β in the above expression one obtains

$$\operatorname{var}(X) = \frac{1}{4(2\beta + 1)},$$

showing that for $\alpha = \beta$ the variance decreases monotonically as Template:Nowwrap increases. Setting $\alpha = \beta = 0$ in this expression, one finds the maximum variance var(X) = 1/4[7] which only occurs approaching the limit, at $\alpha = \beta = 0$.

The beta distribution may also be parametrized in terms of its mean μ ($0 < \mu < 1$) and sample size $\nu = \alpha + \beta$ ($\nu > 0$) (see section below titled "Mean and sample size"):

$$\alpha = \mu\nu, \text{ where } \nu = (\alpha + \beta) > 0$$
$$\beta = (1 - \mu)\nu, \text{ where } \nu = (\alpha + \beta) > 0.$$

Using this parametrization, one can express the variance in terms of the mean μ and the sample size ν as follows:

$$\operatorname{var}(X) = \frac{\mu(1-\mu)}{1+\nu}$$

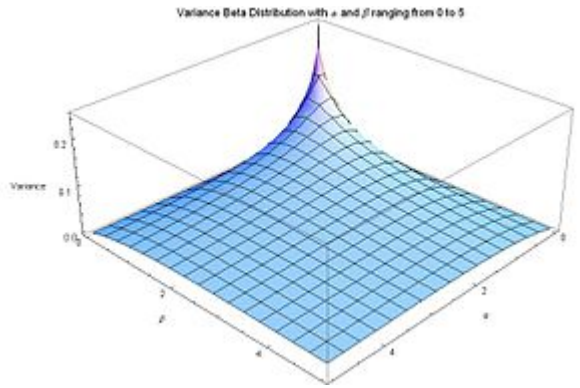
Since {{{1}}}, it must follow that var(X) < $\mu(1 - \mu)$

For a symmetric distribution, the mean is at the middle of the distribution, {{{1}}}, and therefore:

$$\operatorname{var}(X) = \frac{1}{4(1+\nu)} \text{ if } \mu = \tfrac{1}{2}$$

Also, the following limits (with only the noted variable approaching the limit) can be obtained from the above expressions:

$$\lim_{\beta \to 0} \operatorname{var}(X) = \lim_{\alpha \to 0} \operatorname{var}(X) = \lim_{\beta \to \infty} \operatorname{var}(X) = \lim_{\alpha \to \infty} \operatorname{var}(X) = \lim_{\nu \to \infty} \operatorname{var}(X) = \lim_{\mu \to 0} \operatorname{var}(X) = \lim_{\mu \to 1} \operatorname{var}(X) = 0$$
$$\lim_{\nu \to 0} \operatorname{var}(X) = \mu(1 - \mu)$$



Geometric variance and covariance

The logarithm of the geometric variance, ln(var$_{GX}$), of a distribution with random variable X is the second moment of the logarithm of X centered on the geometric mean of X, (ln(G_X):

$$\begin{aligned}\ln \operatorname{var}_{GX} &= \mathrm{E}\left[(\ln X - \ln G_X)^2\right] \\ &= \mathrm{E}[(\ln X - \mathrm{E}\left[\ln X\right])^2] \\ &= \mathrm{E}\left[(\ln X)^2\right] - (\mathrm{E}[\ln X])^2 \\ &= \operatorname{var}[\ln X]\end{aligned}$$

and therefore, the geometric variance is:

$$\operatorname{var}_{GX} = e^{\operatorname{var}[\ln X]}$$

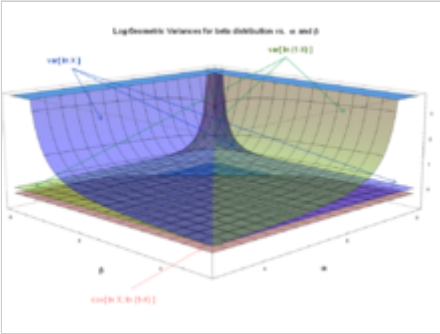
In the Fisher information matrix, and the curvature of the log likelihood function, the logarithm of the geometric variance of the reflected variable (1-X) and the logarithm of the geometric covariance between X and (1-X) appear:

$$\begin{aligned}\ln \operatorname{var}_{G(1\text{-}X)} &= \mathrm{E}[(\ln(1-X) - \ln G_{(1-X)})^2] \\ &= \mathrm{E}[(\ln(1-X) - \mathrm{E}[\ln(1-X)])^2] \\ &= \mathrm{E}[(\ln(1-X))^2] - (\mathrm{E}[\ln(1-X)])^2 \\ &= \operatorname{var}[\ln(1-X)]\end{aligned}$$

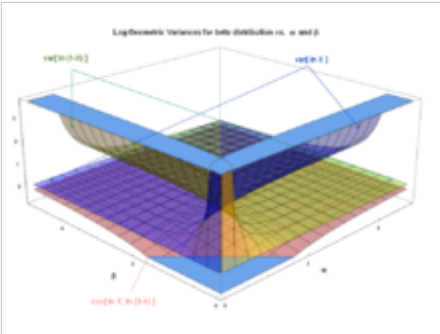
$$\operatorname{var}_{G(1\text{-}X)} = e^{\operatorname{var}[\ln(1-X)]}$$

$$\begin{aligned}\ln \operatorname{cov}_{GX,(1\text{-}X)} &= \mathrm{E}[(\ln X - \ln G_X)(\ln(1-X) - \ln G_{(1-X)})] \\ &= \mathrm{E}[(\ln X - \mathrm{E}[\ln X])(\ln(1-X) - \mathrm{E}[\ln(1-X)])] \\ &= \mathrm{E}[\ln X \ln(1-X)] - \mathrm{E}[\ln X]\,\mathrm{E}[\ln(1-X)] \\ &= \operatorname{cov}[\ln X, \ln(1-X)]\end{aligned}$$

$$\operatorname{cov}_{GX,(1-X)} = e^{\operatorname{cov}[\ln X, \ln(1-X)]}$$



log geometric variances vs. α and β



log geometric variances vs. α and β

For a beta distribution, higher order logarithmic moments can be derived by using the representation of a beta distribution as a proportion of two Gamma distributions and differentiating through the integral. They can be expressed in terms of higher order poly-gamma functions. See the section titled "Other moments, Moments of transformed random variables, Moments of logarithmically transformed random variables". The variance of the logarithmic variables and covariance of ln$X$ and ln(1−X) are:

$$\operatorname{var}[\ln X] = \psi_1(\alpha) - \psi_1(\alpha+\beta)$$
$$\operatorname{var}[\ln(1-X)] = \psi_1(\beta) - \psi_1(\alpha+\beta)$$
$$\operatorname{cov}[\ln X, \ln(1-X)] = -\psi_1(\alpha+\beta)$$

where the **trigamma function**, denoted $\psi_1(\alpha)$, is the second of the polygamma functions, and is defined as the derivative of the digamma function:

$$\psi_1(\alpha) = \frac{d^2 \ln\Gamma(\alpha)}{d\alpha^2} = \frac{d\,\psi(\alpha)}{d\alpha}.$$

Therefore,

$$\ln \operatorname{var}_{GX} = \operatorname{var}[\ln X] = \psi_1(\alpha) - \psi_1(\alpha+\beta)$$
$$\ln \operatorname{var}_{G(1-X)} = \operatorname{var}[\ln(1-X)] = \psi_1(\beta) - \psi_1(\alpha+\beta)$$
$$\ln \operatorname{cov}_{GX,(1-X)} = \operatorname{cov}[\ln X, \ln(1-X)] = -\psi_1(\alpha+\beta)$$

The accompanying plots show the log geometric variances and log geometric covariance versus the shape parameters α and β. The plots show that the log geometric variances and log geometric covariance are close to zero for shape parameters α and β greater than 2, and that the log geometric variances rapidly rise in value for shape parameter values α and β less than unity. The log geometric variances are positive for all values of the shape parameters. The log geometric covariance is negative for all values of the shape parameters, and it reaches large negative values for α and β less than unity.

Following are the limits with one parameter finite (non zero) and the other approaching these limits:

$$\lim_{\alpha\to 0} \ln \operatorname{var}_{GX} = \lim_{\beta\to 0} \ln \operatorname{var}_{G(1-X)} = \infty$$
$$\lim_{\beta\to 0} \ln \operatorname{var}_{GX} = \lim_{\alpha\to\infty} \ln \operatorname{var}_{GX} = \lim_{\alpha\to 0} \ln \operatorname{var}_{G(1-X)} = \lim_{\beta\to\infty} \ln \operatorname{var}_{G(1-X)} = \lim_{\alpha\to\infty} \ln \operatorname{cov}_{GX,(1-X)} = \lim_{\beta\to\infty} \ln \operatorname{cov}_{GX,(1-X)} = 0$$
$$\lim_{\beta\to\infty} \ln \operatorname{var}_{GX} = \psi_1(\alpha)$$
$$\lim_{\alpha\to\infty} \ln \operatorname{var}_{G(1-X)} = \psi_1(\beta)$$
$$\lim_{\alpha\to 0} \ln \operatorname{cov}_{GX,(1-X)} = -\psi_1(\beta)$$
$$\lim_{\beta\to 0} \ln \operatorname{cov}_{GX,(1-X)} = -\psi_1(\alpha)$$

Limits with two parameters varying:

$$\lim_{\alpha\to\infty}\left(\lim_{\beta\to\infty}\ln \operatorname{var}_{GX}\right) = \lim_{\beta\to\infty}\left(\lim_{\alpha\to\infty}\ln \operatorname{var}_{G(1-X)}\right) = \lim_{\alpha\to\infty}\left(\lim_{\beta\to 0}\ln \operatorname{cov}_{GX,(1-X)}\right) = \lim_{\beta\to\infty}\left(\lim_{\alpha\to 0}\ln \operatorname{cov}_{GX,(1-X)}\right) = 0$$

$$\lim_{\alpha\to\infty}\left(\lim_{\beta\to 0}\ln \operatorname{var}_{GX}\right) = \lim_{\beta\to\infty}\left(\lim_{\alpha\to 0}\ln \operatorname{var}_{G(1-X)}\right) = \infty$$

$$\lim_{\alpha\to 0}\left(\lim_{\beta\to 0}\ln \operatorname{cov}_{GX,(1-X)}\right) = \lim_{\beta\to 0}\left(\lim_{\alpha\to 0}\ln \operatorname{cov}_{GX,(1-X)}\right) = -\infty$$

Although both $\ln(\operatorname{var}_{GX})$ and $\ln(\operatorname{var}_{G(1-X)})$ are asymmetric, when the shape parameters are equal, $\alpha = \beta$, one has: $\ln(\operatorname{var}_{GX}) = \ln(\operatorname{var}_{G(1-X)})$. This equality follows from the following symmetry displayed between both log geometric variances:

$$\ln \operatorname{var}_{GX}(B(\alpha,\beta)) = \ln \operatorname{var}_{G(1-X)}(B(\beta,\alpha)).$$

The log geometric covariance is symmetric:

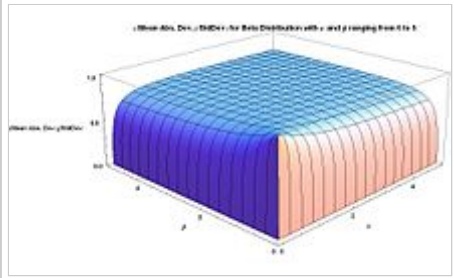
$$\ln \operatorname{cov}_{GX,(1-X)}(B(\alpha,\beta)) = \ln \operatorname{cov}_{GX,(1-X)}(B(\beta,\alpha))$$

Mean absolute deviation around the mean

The mean absolute deviation around the mean for the beta distribution with shape parameters α and β is:[12]

$$\operatorname{E}[|X - E[X]|] = \frac{2\alpha^\alpha \beta^\beta}{B(\alpha,\beta)(\alpha+\beta)^{\alpha+\beta+1}}$$

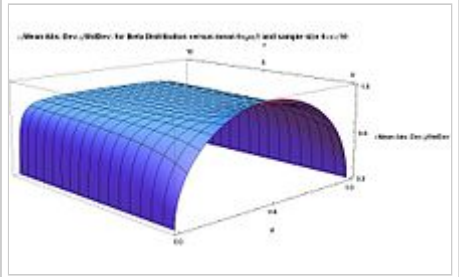
The mean absolute deviation around the mean is a more robust estimator of statistical dispersion than the standard deviation for beta distributions with tails and inflection points at each side of the mode, Beta(α, β) distributions with α,β > 2, as it depends on the linear (absolute) deviations rather than the square deviations from the mean. Therefore, the effect of very large deviations from the mean are not as overly weighted.



Ratio of Mean Abs.Dev. to Std.Dev. for Beta distribution with α and β ranging from 0 to 5

Using Stirling's approximation to the Gamma function, N.L.Johnson and S.Kotz[7] derived the following approximation for values of the shape parameters greater than unity (the relative error for this approximation is only −3.5% for α = β = 1, and it decreases to zero as α → ∞, β → ∞):

$$\frac{\text{mean abs. dev. from mean}}{\text{standard deviation}} = \frac{\mathrm{E}[|X - E[X]|]}{\sqrt{\mathrm{var}(X)}}$$
$$\approx \sqrt{\frac{2}{\pi}}\left(1 + \frac{7}{12(\alpha+\beta)} - \frac{1}{12\alpha} - \frac{1}{12\beta}\right), \text{ if } \alpha, \beta > 1.$$



Ratio of Mean Abs.Dev. to Std.Dev. for Beta distribution with mean $0 \le \mu \le 1$ and sample size $0 < \nu \le 10$

At the limit $\alpha \to \infty$, $\beta \to \infty$, the ratio of the mean absolute deviation to the standard deviation (for the beta distribution) becomes equal to the ratio of the same measures for the normal distribution: $\sqrt{\frac{2}{\pi}}$. For $\alpha = \beta = 1$ this ratio equals $\frac{\sqrt{3}}{2}$, so that from $\alpha = \beta = 1$ to $\alpha, \beta \to \infty$ the ratio decreases by 8.5%. For $\alpha = \beta = 0$ the standard deviation is exactly equal to the mean absolute deviation around the mean. Therefore, this ratio decreases by 15% from $\alpha = \beta = 0$ to $\alpha = \beta = 1$, and by 25% from $\alpha = \beta = 0$ to $\alpha, \beta \to \infty$. However, for skewed beta distributions such that $\alpha \to 0$ or $\beta \to 0$, the ratio of the standard deviation to the mean absolute deviation approaches infinity (although each of them, individually, approaches zero) because the mean absolute deviation approaches zero faster than the standard deviation.

Using the parametrization in terms of mean μ and sample size $\nu = \alpha + \beta > 0$:

$\alpha = \mu\nu$, $\beta = (1-\mu)\nu$

one can express the mean absolute deviation around the mean in terms of the mean μ and the sample size ν as follows:

$$\mathrm{E}[|X - E[X]|] = \frac{2\mu^{\mu\nu}(1-\mu)^{(1-\mu)\nu}}{\nu \mathrm{B}(\mu\nu, (1-\mu)\nu)}$$

For a symmetric distribution, the mean is at the middle of the distribution, $\mu = 1/2$, and therefore:

$$\mathrm{E}[|X - E[X]|] = \frac{2^{1-\nu}}{\nu \mathrm{B}(\frac{\nu}{2}, \frac{\nu}{2})} = \frac{2^{1-\nu}\Gamma(\nu)}{\nu(\Gamma(\frac{\nu}{2}))^2}$$
$$\lim_{\nu \to 0}\left(\lim_{\mu \to \frac{1}{2}} \mathrm{E}[|X - E[X]|]\right) = \tfrac{1}{2}$$
$$\lim_{\nu \to \infty}\left(\lim_{\mu \to \frac{1}{2}} \mathrm{E}[|X - E[X]|]\right) = 0$$

Also, the following limits (with only the noted variable approaching the limit) can be obtained from the above expressions:

$$\lim_{\beta\to 0} \mathrm{E}[|X - E[X]|] = \lim_{\alpha\to 0} \mathrm{E}[|X - E[X]|] = 0$$
$$\lim_{\beta\to \infty} \mathrm{E}[|X - E[X]|] = \lim_{\alpha\to \infty} \mathrm{E}[|X - E[X]|] = 0$$
$$\lim_{\mu\to 0} \mathrm{E}[|X - E[X]|] = \lim_{\mu\to 1} \mathrm{E}[|X - E[X]|] = 0$$
$$\lim_{\nu\to 0} \mathrm{E}[|X - E[X]|] = \sqrt{\mu(1-\mu)}$$
$$\lim_{\nu\to \infty} \mathrm{E}[|X - E[X]|] = 0$$

Mean absolute difference

The mean absolute difference for the Beta distribution is:

$$\mathrm{MD} = \int_0^1 \int_0^1 f(x;\alpha,\beta)\, f(y;\alpha,\beta)\, |x-y|\, dx\, dy = \left(\frac{4}{\alpha+\beta}\right) \frac{B(\alpha+\beta,\alpha+\beta)}{B(\alpha,\alpha)B(\beta,\beta)}$$

The Gini coefficient for the Beta distribution is half of the relative mean absolute difference:

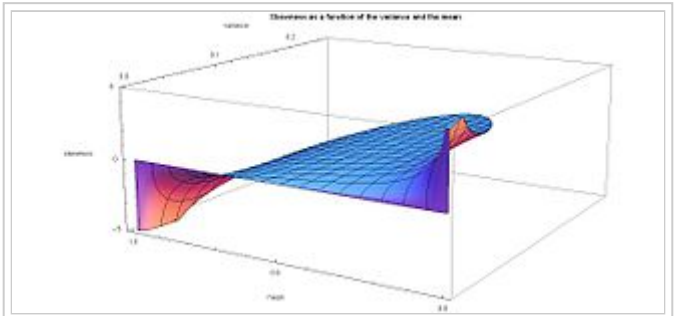
$$\mathrm{G} = \left(\frac{2}{\alpha}\right) \frac{B(\alpha+\beta,\alpha+\beta)}{B(\alpha,\alpha)B(\beta,\beta)}$$

Skewness

The skewness (the third moment centered on the mean, normalized by the 3/2 power of the variance) of the beta distribution is[7]

$$\gamma_1 = \frac{\mathrm{E}[(X-\mu)^3]}{(\mathrm{var}(X))^{3/2}} = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}.$$

Letting α = β in the above expression one obtains $\gamma_1 = 0$, showing once again that for α = β the distribution is symmetric and hence the skewness is zero. Positive skew (right-tailed) for α < β, negative skew (left-tailed) for α > β.



Skewness for Beta Distribution as a function of variance and mean

Using the parametrization in terms of mean μ and sample size ν = α + β:

$$\alpha = \mu\nu, \text{ where } \nu = (\alpha+\beta) > 0$$
$$\beta = (1-\mu)\nu, \text{ where } \nu = (\alpha+\beta) > 0.$$

one can express the skewness in terms of the mean μ and the sample size ν as follows:

$$\gamma_1 = \frac{\mathrm{E}[(X-\mu)^3]}{(\mathrm{var}(X))^{3/2}} = \frac{2(1-2\mu)\sqrt{1+\nu}}{(2+\nu)\sqrt{\mu(1-\mu)}}.$$

The skewness can also be expressed just in terms of the variance *var* and the mean μ as follows:

$$\gamma_1 = \frac{\mathrm{E}[(X-\mu)^3]}{(\mathrm{var}(X))^{3/2}} = \frac{2(1-2\mu)\sqrt{\mathrm{var}}}{\mu(1-\mu)+\mathrm{var}} \text{ if } \mathrm{var} < \mu(1-\mu)$$

The accompanying plot of skewness as a function of variance and mean shows that maximum variance (1/4) is coupled with zero skewness and the symmetry condition (μ = 1/2), and that maximum skewness (positive or negative infinity) occurs when the mean is located at one end or the other, so that the "mass" of the probability distribution is concentrated at the ends (minimum variance).

The following expression for the square of the skewness, in terms of the sample size ν = α + β and the variance *var*, is useful for the method of moments estimation of four parameters:

$$(\gamma_1)^2 = \frac{(\mathrm{E}[(X-\mu)^3])^2}{(\mathrm{var}(X))^3} = \frac{4}{(2+\nu)^2}\left(\frac{1}{\mathrm{var}} - 4(1+\nu)\right)$$

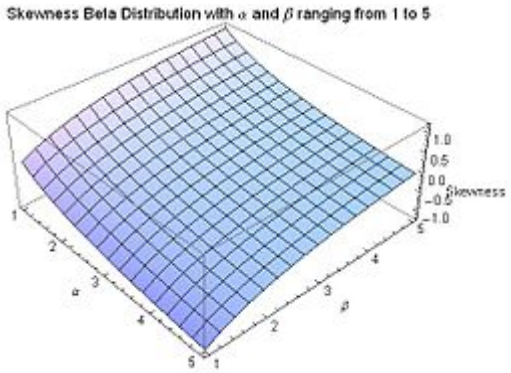
This expression correctly gives a skewness of zero for α = β, since in that case (see section titled "Variance"): $\mathrm{var} = \frac{1}{4(1+\nu)}$.

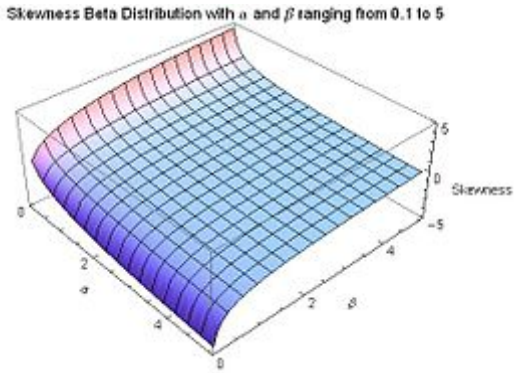
For the symmetric case (α = β), skewness = 0 over the whole range, and the following limits apply:

$$\lim_{\alpha=\beta\to 0} \gamma_1 = \lim_{\alpha=\beta\to\infty} \gamma_1 = \lim_{\nu\to 0} \gamma_1 = \lim_{\nu\to\infty} \gamma_1 = \lim_{\mu\to\frac{1}{2}} \gamma_1 = 0$$

For the asymmetric cases (α ≠ β) the following limits (with only the noted variable approaching the limit) can be obtained from the above expressions:

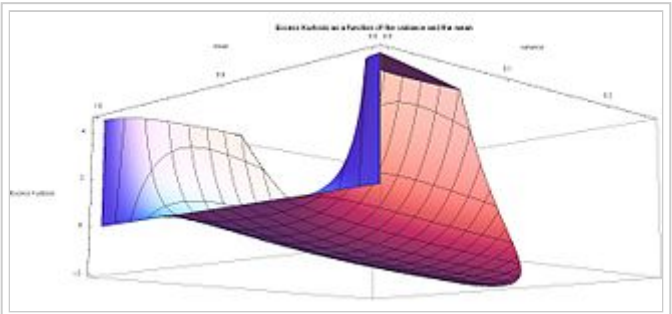
$$\lim_{\alpha\to 0}\gamma_1 = \lim_{\mu\to 0}\gamma_1 = \infty$$
$$\lim_{\beta\to 0}\gamma_1 = \lim_{\mu\to 1}\gamma_1 = -\infty$$
$$\lim_{\alpha\to\infty}\gamma_1 = -\frac{2}{\beta}, \quad \lim_{\beta\to 0}(\lim_{\alpha\to\infty}\gamma_1) = -\infty, \quad \lim_{\beta\to\infty}(\lim_{\alpha\to\infty}\gamma_1) = 0$$
$$\lim_{\beta\to\infty}\gamma_1 = \frac{2}{\alpha}, \quad \lim_{\alpha\to 0}(\lim_{\beta\to\infty}\gamma_1) = \infty, \quad \lim_{\alpha\to\infty}(\lim_{\beta\to\infty}\gamma_1) = 0$$
$$\lim_{\nu\to 0}\gamma_1 = \frac{1-2\mu}{\sqrt{\mu(1-\mu)}}, \quad \lim_{\mu\to 0}(\lim_{\nu\to 0}\gamma_1) = \infty, \quad \lim_{\mu\to 1}(\lim_{\nu\to 0}\gamma_1) = -\infty$$





Kurtosis

The beta distribution has been applied in acoustic analysis to assess damage to gears, as the kurtosis of the beta distribution has been reported to be a good indicator of the condition of a gear.[22] Kurtosis has also been used to distinguish the seismic signal generated by a person's footsteps from other signals. As persons or other targets moving on the ground generate continuous signals in the form of seismic waves, one can separate different targets based on the seismic waves they generate. Kurtosis is sensitive to impulsive signals, so it's much more sensitive to the signal generated by human footsteps than other signals generated by vehicles, winds, noise, etc.[23] Unfortunately, the notation for kurtosis has not been standardized. Kenney and Keeping[24] use the symbol γ_2 for the excess kurtosis, but Abramowitz and Stegun[25] use different terminology. To prevent confusion[26] between kurtosis (the fourth moment centered on the mean, normalized by the square of the variance) and excess kurtosis, when using symbols, they will be spelled out as follows:[12][13]



Excess Kurtosis for Beta Distribution as a function of variance and mean

$$\begin{aligned}\text{excess kurtosis} &= \text{kurtosis} - 3 \\ &= \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{var}(X))^2} - 3 \\ &= \frac{6[\alpha^3 - \alpha^2(2\beta-1) + \beta^2(\beta+1) - 2\alpha\beta(\beta+2)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)} \\ &= \frac{6[(\alpha-\beta)^2(\alpha+\beta+1) - \alpha\beta(\alpha+\beta+2)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}.\end{aligned}$$

Letting α = β in the above expression one obtains

$$\text{excess kurtosis} = -\frac{6}{3+2\alpha} \text{ if } \alpha = \beta.$$

Therefore, for symmetric beta distributions, the excess kurtosis is negative, increasing from a minimum value of −2 at the limit as {α = β} → 0, and approaching a maximum value of zero as {α = β} → ∞. The value of −2 is the minimum value of excess kurtosis that any distribution (not just beta distributions, but any distribution of any possible kind) can ever achieve. This minimum value is reached when all the probability density is entirely concentrated at each end *x* = 0 and *x* = 1, with nothing in between: a 2-point Bernoulli distribution with equal probability 1/2 at each end (a coin toss: see section below "Kurtosis bounded by the square of the skewness" for further discussion). The description of kurtosis as a measure of the "peakedness" (or "heavy tails") of the probability distribution, is strictly applicable to unimodal distributions (for example the normal distribution). However, for more general distributions, like the beta distribution, a more general description of kurtosis is that it is a measure of the proportion of the mass density near the mean. The higher the proportion of mass density near the mean, the higher the kurtosis, while the higher the mass density away from the mean, the lower the kurtosis. For α ≠ β, skewed beta distributions, the excess kurtosis can reach unlimited positive values (particularly for α → 0 for finite β, or for β → 0 for finite α) because all the mass density is concentrated at the mean when the mean coincides with one of the ends. Minimum kurtosis takes place when the mass density is concentrated equally at each end (and therefore the mean is at the center), and there is no probability mass density in between the ends.

Using the parametrization in terms of mean μ and sample size ν = α + β:

$$\begin{aligned}&\alpha = \mu\nu, \text{ where } \nu = (\alpha+\beta) > 0 \\ &\beta = (1-\mu)\nu, \text{ where } \nu = (\alpha+\beta) > 0.\end{aligned}$$

one can express the excess kurtosis in terms of the mean μ and the sample size ν as follows:

$$\text{excess kurtosis} = \frac{6}{3+\nu}\left(\frac{(1-2\mu)^2(1+\nu)}{\mu(1-\mu)(2+\nu)} - 1\right)$$

The excess kurtosis can also be expressed in terms of just the following two parameters: the variance *var*, and the sample size ν as follows:

$$\text{excess kurtosis} = \frac{6}{(3+\nu)(2+\nu)}\left(\frac{1}{\text{var}} - 6 - 5\nu\right) \text{ if } \text{var} < \mu(1-\mu)$$

and, in terms of the variance *var* and the mean μ as follows:

$$\text{excess kurtosis} = \frac{6\,\text{var}\,(1 - \text{var} - 5\mu(1-\mu))}{(\text{var} + \mu(1-\mu))(2\,\text{var} + \mu(1-\mu))} \text{ if } \text{var} < \mu(1-\mu)$$

The plot of excess kurtosis as a function of the variance and the mean shows that the minimum value of the excess kurtosis (−2, which is the minimum possible value for excess kurtosis for any distribution) is intimately coupled with the maximum value of variance (1/4) and the symmetry condition: the mean occurring at the midpoint (μ = 1/2). This occurs for the symmetric case of α = β = 0, with zero skewness. At the limit, this is the 2 point Bernoulli distribution with equal probability 1/2 at each Dirac delta function end *x* = 0 and *x* = 1 and zero probability everywhere else. (A coin toss: one face of the coin being *x* = 0 and the other face being *x* = 1.) Variance is maximum because the distribution is bimodal with nothing in between the two modes (spikes) at each end. Excess kurtosis is minimum: the probability density "mass" is zero at the mean and it is concentrated at the two peaks at each end. Excess kurtosis reaches the minimum possible value (for any distribution) when the probability density function has two spikes at each end: it is bi-"peaky" with nothing in between them.

On the other hand, the plot shows that for extreme skewed cases, where the mean is located near one or the other end (μ = 0 or μ = 1), the variance is close to zero, and the excess kurtosis rapidly approaches infinity when the mean of the distribution approaches either end.

Alternatively, the excess kurtosis can also be expressed in terms of just the following two parameters: the square of the skewness, and the sample size ν as follows:

$$\text{excess kurtosis} = \frac{6}{3+\nu}\left(\frac{(2+\nu)}{4}(\text{skewness})^2 - 1\right) \text{ if } (\text{skewness})^2 - 2 < \text{excess kurtosis} < \frac{3}{2}(\text{skewness})^2$$

From this last expression, one can obtain the same limits published practically a century ago by Karl Pearson in his paper,[27] for the beta distribution (see section below titled "Kurtosis bounded by the square of the skewness"). Setting α + β= ν = 0 in the above expression, one obtains Pearson's lower boundary (values for the skewness and excess kurtosis below the boundary (excess kurtosis + 2 − skewness^2 = 0) cannot occur for any distribution, and hence Karl Pearson appropriately called the region below this boundary the "impossible region"). The limit of α + β = ν → ∞ determines Pearson's upper boundary.

$$\lim_{\nu\to 0} \text{excess kurtosis} = (\text{skewness})^2 - 2$$
$$\lim_{\nu\to \infty} \text{excess kurtosis} = \tfrac{3}{2}(\text{skewness})^2$$

therefore:

$$(\text{skewness})^2 - 2 < \text{excess kurtosis} < \tfrac{3}{2}(\text{skewness})^2$$

Values of $\nu = \alpha + \beta$ such that ν ranges from zero to infinity, $0 < \nu < \infty$, span the whole region of the beta distribution in the plane of excess kurtosis versus squared skewness.

For the symmetric case ($\alpha = \beta$), the following limits apply:

$$\lim_{\alpha=\beta\to 0} \text{excess kurtosis} = -2$$

$$\lim_{\alpha=\beta\to \infty} \text{excess kurtosis} = 0$$

$$\lim_{\mu\to \frac{1}{2}} \text{excess kurtosis} = -\frac{6}{3+\nu}$$

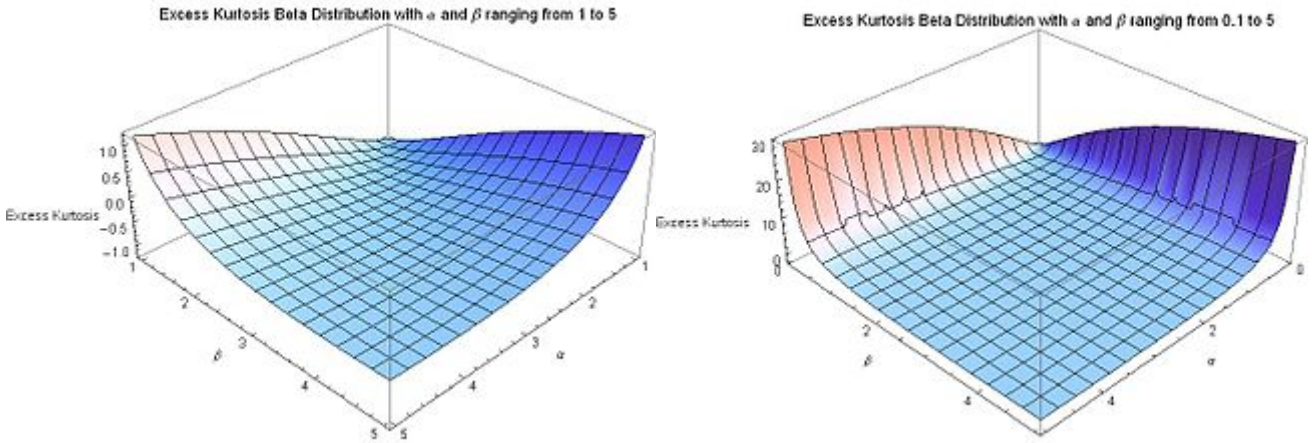
For the unsymmetric cases ($\alpha \neq \beta$) the following limits (with only the noted variable approaching the limit) can be obtained from the above expressions:

$$\lim_{\alpha\to 0} \text{excess kurtosis} = \lim_{\beta\to 0} \text{excess kurtosis} = \lim_{\mu\to 0} \text{excess kurtosis} = \lim_{\mu\to 1} \text{excess kurtosis} = \infty$$

$$\lim_{\alpha\to \infty} \text{excess kurtosis} = \frac{6}{\beta},\ \lim_{\beta\to 0}\left(\lim_{\alpha\to \infty} \text{excess kurtosis}\right) = \infty,\ \lim_{\beta\to \infty}\left(\lim_{\alpha\to \infty} \text{excess kurtosis}\right) = 0$$

$$\lim_{\beta\to \infty} \text{excess kurtosis} = \frac{6}{\alpha},\ \lim_{\alpha\to 0}\left(\lim_{\beta\to \infty} \text{excess kurtosis}\right) = \infty,\ \lim_{\alpha\to \infty}\left(\lim_{\beta\to \infty} \text{excess kurtosis}\right) = 0$$

$$\lim_{\nu\to 0} \text{excess kurtosis} = -6 + \frac{1}{\mu(1-\mu)},\ \lim_{\mu\to 0}\left(\lim_{\nu\to 0} \text{excess kurtosis}\right) = \infty,\ \lim_{\mu\to 1}\left(\lim_{\nu\to 0} \text{excess kurtosis}\right) = \infty$$



Characteristic function

The characteristic function is the Fourier transform of the probability density function. The characteristic function of the beta distribution is Kummer's confluent hypergeometric function (of the first kind):[7][25][28]

$$\begin{aligned}\varphi_X(\alpha;\beta;t) &= \mathbf{E}\left[e^{itX}\right] \\ &= \int_0^1 e^{itx} f(x;\alpha,\beta)dx \\ &= {}_1F_1(\alpha;\alpha+\beta;it) \\ &= \sum_{n=0}^{\infty} \frac{\alpha^{(n)}(it)^n}{(\alpha+\beta)^{(n)}n!} \\ &= 1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{(it)^k}{k!}\end{aligned}$$

where

$$x^{(n)} = x(x+1)(x+2)\cdots(x+n-1)$$

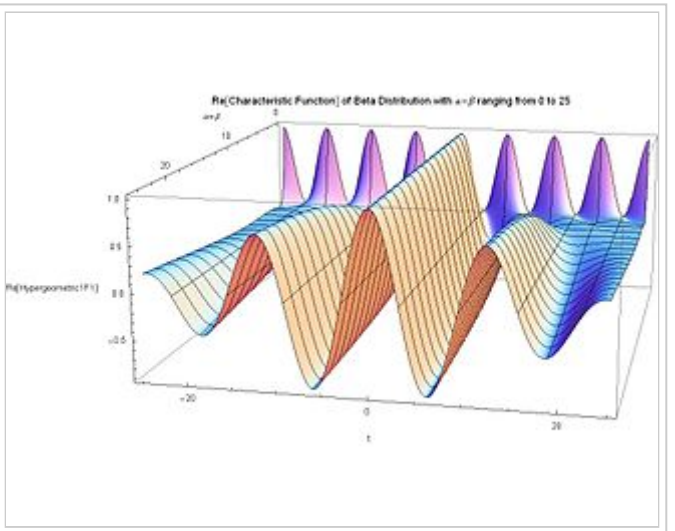
is the rising factorial, also called the "Pochhammer symbol". The value of the characteristic function for $t = 0$, is one:

$$\varphi_X(\alpha;\beta;0) = {}_1F_1(\alpha;\alpha+\beta;0) = 1\,.$$

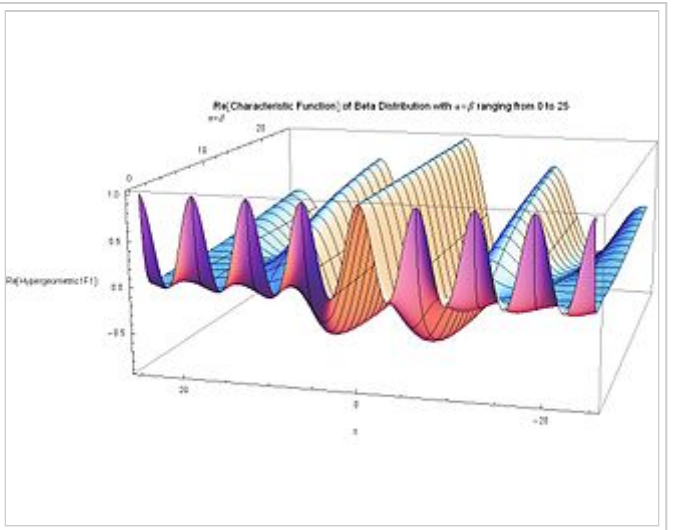
Also, the real and imaginary parts of the characteristic function enjoy the following symmetries with respect to the origin of variable t:

$$\begin{aligned}\mathbf{Re}\left[{}_1F_1(\alpha;\alpha+\beta;it)\right] &= \mathbf{Re}\left[{}_1F_1(\alpha;\alpha+\beta;-it)\right] \\ \mathbf{Im}\left[{}_1F_1(\alpha;\alpha+\beta;it)\right] &= -\mathbf{Im}\left[{}_1F_1(\alpha;\alpha+\beta;-it)\right]\end{aligned}$$

The symmetric case α = β simplifies the characteristic function of the beta distribution to a Bessel function, since in the special case α + β = 2α the confluent hypergeometric function (of the first kind) reduces to a Bessel function (the modified Bessel function of the first kind $I_{\alpha-\frac{1}{2}}$) using Kummer's second transformation as follows:



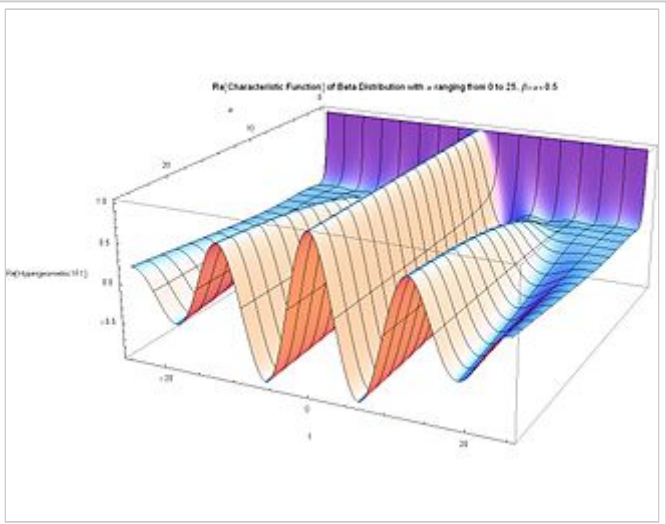
Re(characteristic function) symmetric case α = β ranging from 25 to 0



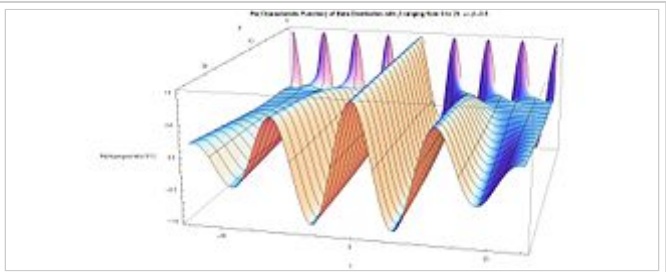
Re(characteristic function) symmetric case α = β ranging from 0 to 25

$$
\begin{aligned}
{}_1F_1(\alpha;2\alpha;it) &= e^{\frac{it}{2}}\,{}_0F_1\left(;\alpha+\tfrac{1}{2};\frac{(it)^2}{16}\right) \\
&= e^{\frac{it}{2}}\left(\frac{it}{4}\right)^{\frac{1}{2}-\alpha}\Gamma\left(\alpha+\tfrac{1}{2}\right)I_{\alpha-\frac{1}{2}}\left(\frac{it}{2}\right).
\end{aligned}
$$

In the accompanying plots, the real part (Re) of the characteristic function of the beta distribution is displayed for symmetric (α = β) and skewed (α ≠ β) cases.



Re(characteristic function) β = α + 1/2; α ranging from 25 to 0



Re(characteristic function) α = β + 1/2; β ranging from 25 to 0

Other moments

Moment generating function

It also follows[7][12] that the moment generating function is

$$
\begin{aligned}
M_X(\alpha;\beta;t) &= \mathrm{E}\left[e^{tX}\right] \\
&= \int_0^1 e^{tx} f(x;\alpha,\beta)\,dx \\
&= {}_1F_1(\alpha;\alpha+\beta;t) \\
&= \sum_{n=0}^{\infty}\frac{\alpha^{(n)}}{(\alpha+\beta)^{(n)}}\frac{t^n}{n!} \\
&= 1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}
\end{aligned}
$$

In particular $M_X(\alpha;\beta;0) = 1$.

Higher moments

Using the moment generating function, the k-th raw moment is given by[7] the factor

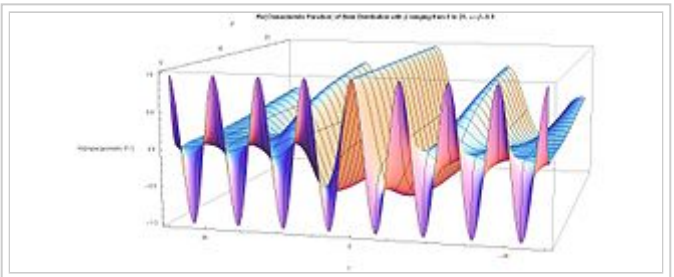
$$
\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}
$$

multiplying the (exponential series) term $\left(\frac{t^k}{k!}\right)$ in the series of the moment generating function

$$\mathbf{E}[X^k] = \frac{\alpha^{(k)}}{(\alpha+\beta)^{(k)}} = \prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r}$$

where $(x)^{(k)}$ is a Pochhammer symbol representing rising factorial. It can also be written in a recursive form as

$$\mathbf{E}[X^k] = \frac{\alpha+k-1}{\alpha+\beta+k-1}\,\mathbf{E}[X^{k-1}].$$



Re(characteristic function) α = β + 1/2; β ranging from 0 to 25

Since the moment generating function $M_X(\alpha;\beta;\cdot)$ has a positive radius of convergence, the beta distribution is determined by its moments.[29]

Moments of transformed random variables

Moments of linearly transformed, product and inverted random variables

One can also show the following expectations for a transformed random variable,[7] where the random variable X is Beta-distributed with parameters α and β: $X \sim$ Beta(α, β). The expected value of the variable $(1-X)$ is the mirror-symmetry of the expected value based on X:

$$\mathbf{E}[1-X] = \frac{\beta}{\alpha+\beta}$$

$$\mathbf{E}[X(1-X)] = \mathbf{E}[(1-X)X] = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}$$

Due to the mirror-symmetry of the probability density function of the beta distribution, the variances based on variables X and $(1-X)$ are identical, and the covariance on $X(1-X)$ is the negative of the variance:

$$\mathbf{var}[(1-X)] = \mathbf{var}[X] = -\mathbf{cov}[X,(1-X)] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

These are the expected values for inverted variables, (these are related to the harmonic means, see section titled "Harmonic mean"):

$$\mathbf{E}\left[\frac{1}{X}\right] = \frac{\alpha+\beta-1}{\alpha-1} \text{ if } \alpha > 1$$
$$\mathbf{E}\left[\frac{1}{1-X}\right] = \frac{\alpha+\beta-1}{\beta-1} \text{ if } \beta > 1$$

The following transformation by dividing the variable X by its mirror-image $X/(1-X)$ results in the expected value of the "inverted beta distribution" or beta prime distribution (also known as beta distribution of the second kind or Pearson's Type VI):[7]

$$\mathbf{E}\left[\frac{X}{1-X}\right] = \frac{\alpha}{\beta-1} \text{ if } \beta > 1$$
$$\mathbf{E}\left[\frac{1-X}{X}\right] = \frac{\beta}{\alpha-1} \text{ if } \alpha > 1$$

Variances of these transformed variables can be obtained by integration, as the expected values of the second moments centered on the corresponding variables:

$$\mathbf{var}\left[\frac{1}{X}\right] = \mathbf{E}\left[\left(\frac{1}{X} - \mathbf{E}\left[\frac{1}{X}\right]\right)^2\right] =$$
$$\mathbf{var}\left[\frac{1-X}{X}\right] = \mathbf{E}\left[\left(\frac{1-X}{X} - \mathbf{E}\left[\frac{1-X}{X}\right]\right)^2\right] = \frac{\beta(\alpha+\beta-1)}{(\alpha-2)(\alpha-1)^2} \text{ if } \alpha > 2$$

The following variance of the variable X divided by its mirror-image ($X/(1-X)$ results in the variance of the "inverted beta distribution" or beta prime distribution (also known as beta distribution of the second kind or Pearson's Type VI):[7]

$$\mathbf{var}\left[\frac{1}{1-X}\right] = \mathbf{E}\left[\left(\frac{1}{1-X} - \mathbf{E}\left[\frac{1}{1-X}\right]\right)^2\right] = \mathbf{var}\left[\frac{X}{1-X}\right] =$$
$$\mathbf{E}\left[\left(\frac{X}{1-X} - \mathbf{E}\left[\frac{X}{1-X}\right]\right)^2\right] = \frac{\alpha(\alpha+\beta-1)}{(\beta-2)(\beta-1)^2} \text{ if } \beta > 2$$

The covariances are:

$$\mathbf{cov}\left[\frac{1}{X}, \frac{1}{1-X}\right] = \mathbf{cov}\left[\frac{1-X}{X}, \frac{X}{1-X}\right] = \mathbf{cov}\left[\frac{1}{X}, \frac{X}{1-X}\right] = \mathbf{cov}\left[\frac{1-X}{X}, \frac{1}{1-X}\right] = \frac{\alpha+\beta-1}{(\alpha-1)(\beta-1)} \text{ if } \alpha, \beta > 1$$

These expectations and variances appear in the four-parameter Fisher information matrix (section titled "Fisher information," "four parameters")

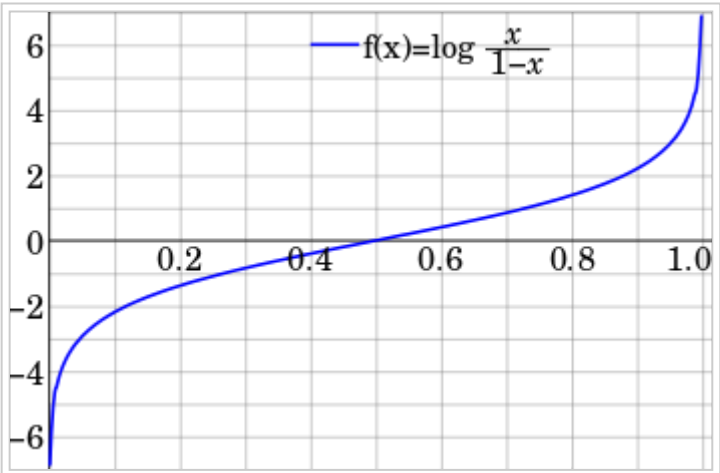
Moments of logarithmically transformed random variables

Expected values for logarithmic transformations (useful for maximum likelihood estimates, see section titled "Parameter estimation, Maximum likelihood" below) are discussed in this section. The following logarithmic linear transformations are related to the geometric means G_X and $G_{(1-X)}$ (see section titled "Geometric mean"):

$$\mathbf{E}[\ln(X)] = \psi(\alpha) - \psi(\alpha+\beta) = -\mathbf{E}\left[\ln\left(\frac{1}{X}\right)\right],$$
$$\mathbf{E}[\ln(1-X)] = \psi(\beta) - \psi(\alpha+\beta) = -\mathbf{E}\left[\ln\left(\frac{1}{1-X}\right)\right].$$

Where the **digamma function** ψ(α) is defined as the logarithmic derivative of the gamma function:[25]

$$\psi(\alpha) = \frac{d\ln\Gamma(\alpha)}{d\alpha}$$



Plot of logit(X) = ln(X/(1−X)) (vertical axis) vs. X in the domain of 0 to 1 (horizontal axis). Logit transformations are interesting, as they usually transform various shapes (including J-shapes) into (usually skewed) bell-shaped densities over the logit variable, and they may remove the end singularities over the original variable

Logit transformations are interesting,[30] as they usually transform various shapes (including J-shapes) into (usually skewed) bell-shaped densities over the logit variable, and they may remove the end singularities over the original variable:

$$\mathbf{E}\left[\ln\left(\frac{X}{1-X}\right)\right] = \psi(\alpha) - \psi(\beta) = \mathbf{E}[\ln(X)] + \mathbf{E}\left[\ln\left(\frac{1}{1-X}\right)\right],$$
$$\mathbf{E}\left[\ln\left(\frac{1-X}{X}\right)\right] = \psi(\beta) - \psi(\alpha) = -\mathbf{E}\left[\ln\left(\frac{X}{1-X}\right)\right].$$

Johnson[31] considered the distribution of the logit - transformed variable ln(X/1−X), including its moment generating function and approximations for large values of the shape parameters. This transformation extends the finite support [0, 1] based on the original variable X to infinite support in both directions of the real line (−∞, +∞).

Higher order logarithmic moments can be derived by using the representation of a beta distribution as a proportion of two Gamma distributions and differentiating through the integral. They can be expressed in terms of higher order poly-gamma functions as follows:

$$\begin{aligned}\mathbf{E}\left[\ln^2(X)\right] &= (\psi(\alpha) - \psi(\alpha+\beta))^2 + \psi_1(\alpha) - \psi_1(\alpha+\beta),\\ \mathbf{E}\left[\ln^2(1-X)\right] &= (\psi(\beta) - \psi(\alpha+\beta))^2 + \psi_1(\beta) - \psi_1(\alpha+\beta),\\ \mathbf{E}[\ln(X)\ln(1-X)] &= (\psi(\alpha) - \psi(\alpha+\beta))(\psi(\beta) - \psi(\alpha+\beta)) - \psi_1(\alpha+\beta).\end{aligned}$$

therefore the variance of the logarithmic variables and covariance of ln(X) and ln(1−X) are:

$$\mathbf{cov}[\ln(X), \ln(1-X)] = \mathbf{E}[\ln(X)\ln(1-X)] - \mathbf{E}[\ln(X)]\,\mathbf{E}[\ln(1-X)] = -\psi_1(\alpha+\beta)$$

$$\begin{aligned}\mathbf{var}[\ln X] &= \mathbf{E}[\ln^2(X)] - (\mathbf{E}[\ln(X)])^2\\ &= \psi_1(\alpha) - \psi_1(\alpha+\beta)\\ &= \psi_1(\alpha) + \mathbf{cov}[\ln(X), \ln(1-X)]\end{aligned}$$

$$\begin{aligned}\mathbf{var}[\ln(1-X)] &= \mathbf{E}[\ln^2(1-X)] - (\mathbf{E}[\ln(1-X)])^2\\ &= \psi_1(\beta) - \psi_1(\alpha+\beta)\\ &= \psi_1(\beta) + \mathbf{cov}[\ln(X), \ln(1-X)]\end{aligned}$$

where the **trigamma function**, denoted $\psi_1(\alpha)$, is the second of the polygamma functions, and is defined as the derivative of the digamma function:

$$\psi_1(\alpha) = \frac{d^2\ln\Gamma(\alpha)}{d\alpha^2} = \frac{d\psi(\alpha)}{d\alpha}.$$

The variances and covariance of the logarithmically transformed variables X and (1−X) are different, in general, because the logarithmic transformation destroys the mirror-symmetry of the original variables X and (1−X), as the logarithm approaches negative infinity for the variable approaching zero.

These logarithmic variances and covariance are the elements of the Fisher information matrix for the beta distribution. They are also a measure of the curvature of the log likelihood function (see section on Maximum likelihood estimation).

The variances of the log inverse variables are identical to the variances of the log variables:

$$\operatorname{var}\left[\ln\left(\frac{1}{X}\right)\right] = \operatorname{var}[\ln(X)] = \psi_1(\alpha) - \psi_1(\alpha+\beta),$$
$$\operatorname{var}\left[\ln\left(\frac{1}{1-X}\right)\right] = \operatorname{var}[\ln(1-X)] = \psi_1(\beta) - \psi_1(\alpha+\beta),$$
$$\operatorname{cov}\left[\ln\left(\frac{1}{X}\right), \ln\left(\frac{1}{1-X}\right)\right] = \operatorname{cov}[\ln(X), \ln(1-X)] = -\psi_1(\alpha+\beta).$$

It also follows that the variances of the logit transformed variables are:

$$\operatorname{var}\left[\ln\left(\frac{X}{1-X}\right)\right] = \operatorname{var}\left[\ln\left(\frac{1-X}{X}\right)\right] = -\operatorname{cov}\left[\ln\left(\frac{X}{1-X}\right), \ln\left(\frac{1-X}{X}\right)\right] = \psi_1(\alpha) + \psi_1(\beta)$$

Quantities of information (entropy)

Given a beta distributed random variable, $X \sim \text{Beta}(\alpha, \beta)$, the differential entropy of X is[32](measured in nats), the expected value of the negative of the logarithm of the probability density function:

$$\begin{aligned} h(X) &= \mathrm{E}[-\ln(f(x;\alpha,\beta))] \\ &= \int_0^1 -f(x;\alpha,\beta)\ln(f(x;\alpha,\beta))\,dx \\ &= \ln(\mathrm{B}(\alpha,\beta)) - (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta) + (\alpha+\beta-2)\psi(\alpha+\beta) \end{aligned}$$

where $f(x; \alpha, \beta)$ is the probability density function of the beta distribution:

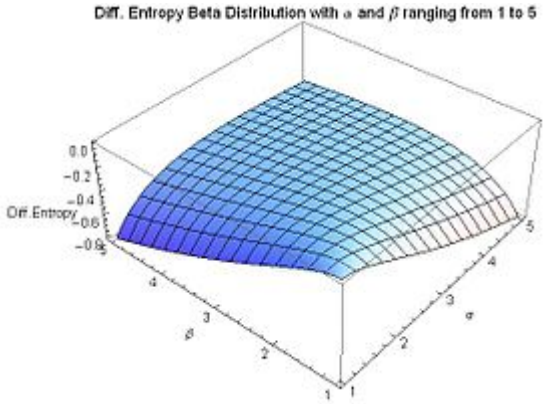
$$f(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

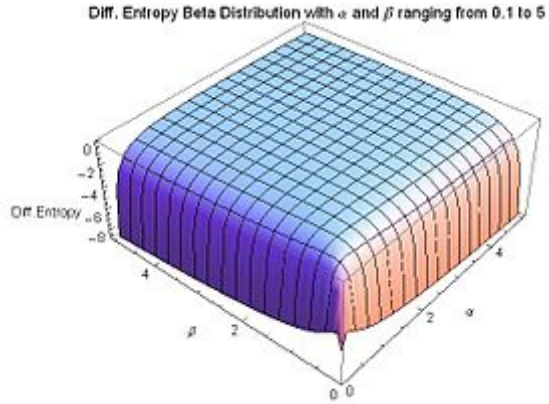
The digamma function ψ appears in the formula for the differential entropy as a consequence of Euler's integral formula for the harmonic numbers which follows from the integral:

$$\int_0^1 \frac{1-x^{\alpha-1}}{1-x}\,dx = \psi(\alpha) - \psi(1)$$

The differential entropy of the beta distribution is negative for all values of α and β greater than zero, except at α = β = 1 (for which values the beta distribution is the same as the uniform distribution), where the differential entropy reaches its maximum value of zero. It is to be expected that the maximum entropy should take place when the beta distribution becomes equal to the uniform distribution, since uncertainty is maximal when all possible events are equiprobable.

For α or β approaching zero, the differential entropy approaches its minimum value of negative infinity. For (either or both) α or β approaching zero, there is a maximum amount of order: all the probability density is concentrated at the ends, and there is zero probability density at points located between the ends. Similarly for (either or both) α or β approaching infinity, the differential entropy approaches its minimum value of negative infinity, and a maximum amount of order. If either α or β approaches infinity (and the other is finite) all the probability density is concentrated at an end, and the probability density is zero everywhere else. If both shape parameters are equal (the symmetric case), α = β, and they approach infinity simultaneously, the probability density becomes a spike (Dirac delta function) concentrated at the middle $x = 1/2$, and hence there is 100% probability at the middle $x = 1/2$ and zero probability everywhere else.





The (continuous case) differential entropy was introduced by Shannon in his original paper (where he named it the "entropy of a continuous distribution"), as the concluding part[33] of the same paper where he defined the discrete entropy. It is known since then that the differential entropy may differ from the infinitesimal limit of the discrete entropy by an infinite offset, therefore the differential entropy can be negative (as it is for the beta distribution). What really matters is the relative value of entropy.

Given two beta distributed random variables, $X_1 \sim \text{Beta}(\alpha, \beta)$ and $X_2 \sim \text{Beta}(\alpha', \beta')$, the cross entropy is (measured in nats)[34]

$$\begin{aligned} H(X_1, X_2) &= \int_0^1 -f(x;\alpha,\beta)\ln(f(x;\alpha',\beta'))dx \\ &= \ln(\mathrm{B}(\alpha',\beta')) - (\alpha'-1)\psi(\alpha) - (\beta'-1)\psi(\beta) + (\alpha'+\beta'-2)\psi(\alpha+\beta). \end{aligned}$$

The cross entropy has been used as an error metric to measure the distance between two hypotheses.[35][36] Its absolute value is minimum when the two distributions are identical. It is the information measure most closely related to the log maximum likelihood [34](see section on "Parameter estimation. Maximum likelihood estimation")).

The relative entropy, or Kullback–Leibler divergence $D_{\mathrm{KL}}(X_1, X_2)$, is a measure of the inefficiency of assuming that the distribution is $X_2 \sim \mathrm{Beta}(\alpha', \beta')$ when the distribution is really $X_1 \sim \mathrm{Beta}(\alpha, \beta)$. It is defined as follows (measured in nats).

$$\begin{aligned} D_{\mathrm{KL}}(X_1, X_2) &= \int_0^1 f(x;\alpha,\beta) \ln\left(\frac{f(x;\alpha,\beta)}{f(x;\alpha',\beta')}\right) dx \\ &= \left(\int_0^1 f(x;\alpha,\beta)\ln(f(x;\alpha,\beta))dx\right) - \left(\int_0^1 f(x;\alpha,\beta)\ln(f(x;\alpha',\beta'))dx\right) \\ &= -h(X_1) + H(X_1, X_2) \\ &= \ln\left(\frac{\mathrm{B}(\alpha',\beta')}{\mathrm{B}(\alpha,\beta)}\right) + (\alpha-\alpha')\psi(\alpha) + (\beta-\beta')\psi(\beta) + (\alpha'-\alpha+\beta'-\beta)\psi(\alpha+\beta). \end{aligned}$$

The relative entropy, or Kullback–Leibler divergence, is always non-negative. A few numerical examples follow:

- $X_1 \sim \mathrm{Beta}(1, 1)$ and $X_2 \sim \mathrm{Beta}(3, 3)$; $D_{\mathrm{KL}}(X_1, X_2) = 0.598803$; $D_{\mathrm{KL}}(X_2, X_1) = 0.267864$; $h(X_1) = 0$; $h(X_2) = -0.267864$
- $X_1 \sim \mathrm{Beta}(3, 0.5)$ and $X_1 \sim \mathrm{Beta}(0.5, 3)$; $D_{\mathrm{KL}}(X_1, X_2) = 7.21574$; $D_{\mathrm{KL}}(X_2, X_1) = 7.21574$; $h(X_1) = -1.10805$; $h(X_2) = -1.10805$.

The Kullback–Leibler divergence is not symmetric $D_{\mathrm{KL}}(X_1, X_2) \neq D_{\mathrm{KL}}(X_2, X_1)$ for the case in which the individual beta distributions Beta(1, 1) and Beta(3, 3) are symmetric, but have different entropies $h(X_1) \neq h(X_2)$. The value of the Kullback divergence depends on the direction traveled: whether going from a higher (differential) entropy to a lower (differential) entropy or the other way around. In the numerical example above, the Kullback divergence measures the inefficiency of assuming that the distribution is (bell-shaped) Beta(3, 3), rather than (uniform) Beta(1, 1). The "h" entropy of Beta(1, 1) is higher than the "h" entropy of Beta(3, 3) because the uniform distribution Beta(1, 1) has a maximum amount of disorder. The Kullback divergence is more than two times higher (0.598803 instead of 0.267864) when measured in the direction of decreasing entropy: the direction that assumes that the (uniform) Beta(1, 1) distribution is (bell-shaped) Beta(3, 3) rather than the other way around. In this restricted sense, the Kullback divergence is consistent with the second law of thermodynamics.

The Kullback–Leibler divergence is symmetric $D_{\mathrm{KL}}(X_1, X_2) = D_{\mathrm{KL}}(X_2, X_1)$ for the skewed cases Beta(3, 0.5) and Beta(0.5, 3) that have equal differential entropy $h(X_1) = h(X_2)$.

The symmetry condition:

$$D_{\mathrm{KL}}(X_1, X_2) = D_{\mathrm{KL}}(X_2, X_1), \text{ if } h(X_1) = h(X_2), \text{ for (skewed) } \alpha \neq \beta$$

follows from the above definitions and the mirror-symmetry $f(x; \alpha, \beta) = f(1-x; \alpha, \beta)$ enjoyed by the beta distribution.

Relationships between statistical measures

Mean, mode and median relationship

If $1 < \alpha < \beta$ then mode ≤ median ≤ mean.[17] Expressing the mode (only for $\alpha, \beta > 1$), and the mean in terms of α and β:

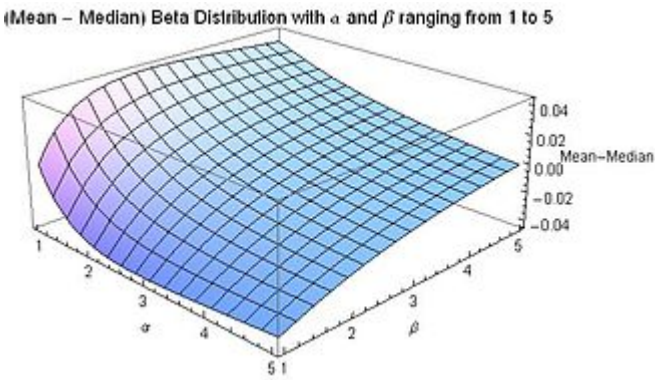
$$\frac{\alpha - 1}{\alpha + \beta - 2} \leq \text{median} \leq \frac{\alpha}{\alpha + \beta},$$

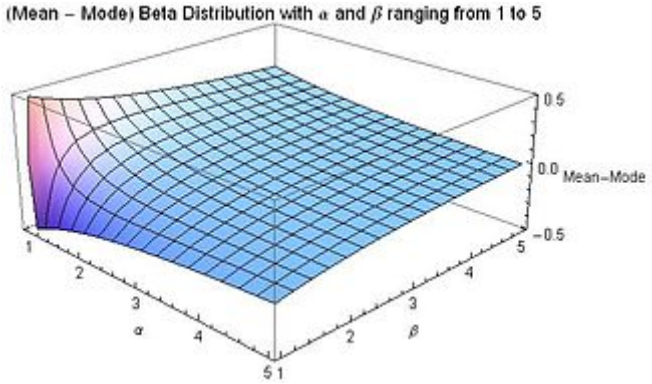
If $1 < \beta < \alpha$ then the order of the inequalities are reversed. For $\alpha, \beta > 1$ the absolute distance between the mean and the median is less than 5% of the distance between the maximum and minimum values of *x*. On the other hand, the absolute distance between the mean and the mode can reach 50% of the distance between the maximum and minimum values of *x*, for the (pathological) case of $\alpha = 1$ and $\beta = 1$ (for which values the beta distribution approaches the uniform distribution and the differential entropy approaches its maximum value, and hence maximum "disorder").

For example, for $\alpha = 1.0001$ and $\beta = 1.00000001$:

- mode = 0.9999; PDF(mode) = 1.00010
- mean = 0.500025; PDF(mean) = 1.00003
- median = 0.500035; PDF(median) = 1.00003
- mean − mode = −0.499875
- mean − median = -9.65538×10^{-6}

(where PDF stands for the value of the probability density function)

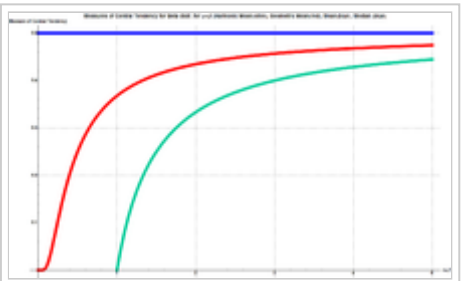




Mean, geometric mean and harmonic mean relationship

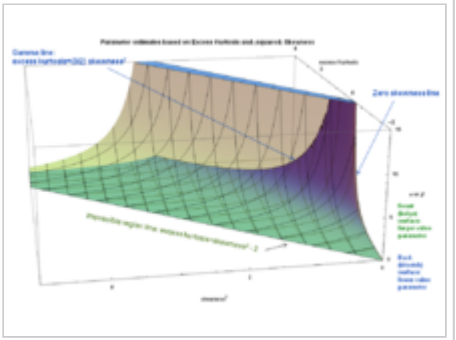
It is known from the inequality of arithmetic and geometric means that the geometric mean is lower than the mean. Similarly, the harmonic mean is lower than the geometric mean. The accompanying plot shows that for $\alpha = \beta$, both the mean and the median are exactly equal to 1/2, regardless of the value of $\alpha = \beta$, and the mode is also equal to 1/2 for $\alpha = \beta > 1$, however the geometric and harmonic means are lower than 1/2 and they only approach this value asymptotically as $\alpha = \beta$

$\to \infty$.



:Mean, Median, Geometric Mean and Harmonic Mean for Beta distribution with $0 < \alpha = \beta < 5$

Kurtosis bounded by the square of the skewness



Beta distribution α and β parameters vs. excess Kurtosis and squared Skewness

As remarked by Feller,[11] in the Pearson system the beta probability density appears as type I (any difference between the beta distribution and Pearson's type I distribution is only superficial and it makes no difference for the following discussion regarding the relationship between kurtosis and skewness). Karl Pearson showed, in Plate 1 of his paper [27] published in 1916, a graph with the kurtosis as the vertical axis (ordinate) and the square of the skewness as the horizontal axis (abscissa), in which a number of distributions were displayed.[37] The region occupied by the beta distribution is bounded by the following two lines in the (skewness^2,kurtosis) plane, or the (skewness^2,excess kurtosis) plane:

$$(\text{skewness})^2 + 1 < \text{kurtosis} < \frac{3}{2}(\text{skewness})^2 + 3$$

or, equivalently,

$$(\text{skewness})^2 - 2 < \text{excess kurtosis} < \frac{3}{2}(\text{skewness})^2$$

(At a time when there were no powerful digital computers), Karl Pearson accurately computed further boundaries,[10][27] for example, separating the "U-shaped" from the "J-shaped" distributions. The lower boundary line (excess kurtosis + 2 − skewness^2 = 0) is produced by skewed "U-shaped" beta distributions with both values of shape parameters α and β close to zero. The upper boundary line (excess kurtosis − (3/2) skewness^2 = 0) is produced by extremely skewed distributions with very large values of one of the parameters and very small values of the other parameter. Karl Pearson showed [27] that this upper boundary line (excess kurtosis − (3/2) skewness^2 = 0) is also the intersection with Pearson's distribution III, which has unlimited support in one direction (towards positive infinity), and can be bell-shaped or J-shaped. His son, Egon Pearson, showed [37] that the region (in the kurtosis/squared-skewness plane) occupied by the beta distribution (equivalently, Pearson's distribution I) as it approaches this boundary (excess kurtosis − (3/2) skewness^2 = 0) is shared with the noncentral chi-squared distribution. Karl Pearson[38] (Pearson 1895, pp. 357, 360, 373–376) also showed that the gamma distribution is a Pearson type III distribution. Hence this boundary line for Pearson's type III distribution is known as the gamma line. (This can be shown from the fact that the excess kurtosis of the gamma distribution is 6/k and the square of the skewness is 4/k, hence (excess kurtosis − (3/2) skewness^2 = 0) is identically satisfied by the gamma distribution regardless of the value of the parameter "k"). Pearson later noted that the chi-squared distribution is a special case of Pearson's type III and also shares this boundary line (as it is apparent from

the fact that for the chi-squared distribution the excess kurtosis is $12/k$ and the square of the skewness is $8/k$, hence (excess kurtosis − (3/2) skewness^2 = 0) is identically satisfied regardless of the value of the parameter "k"). This is to be expected, since the chi-squared distribution $X \sim \chi^2(k)$ is a special case of the gamma distribution, with parametrization X ~ Γ(k/2, 1/2) where k is a positive integer that specifies the "number of degrees of freedom" of the chi-squared distribution.

An example of a beta distribution near the upper boundary (excess kurtosis − (3/2) skewness^2 = 0) is given by α = 0.1, β = 1000, for which the ratio (excess kurtosis)/(skewness^2) = 1.49835 approaches the upper limit of 1.5 from below. An example of a beta distribution near the lower boundary (excess kurtosis + 2 − skewness^2 = 0) is given by α= 0.0001, β = 0.1, for which values the expression (excess kurtosis + 2)/(skewness^2) = 1.01621 approaches the lower limit of 1 from above. In the infinitesimal limit for both α and β approaching zero symmetrically, the excess kurtosis reaches its minimum value at −2. This minimum value occurs at the point at which the lower boundary line intersects the vertical axis (ordinate). (However, in Pearson's original chart, the ordinate is kurtosis, instead of excess kurtosis, and it increases downwards rather than upwards).

Values for the skewness and excess kurtosis below the lower boundary (excess kurtosis + 2 − skewness^2 = 0) cannot occur for any distribution, and hence Karl Pearson appropriately called the region below this boundary the "impossible region." The boundary for this "impossible region" is determined by (symmetric or skewed) bimodal "U"-shaped distributions for which parameters α and β approach zero and hence all the probability density is concentrated at the ends: x = 0, 1 with practically nothing in between them. Since for α ≈ β ≈ 0 the probability density is concentrated at the two ends x = 0 and x = 1, this "impossible boundary" is determined by a 2-point distribution: the probability can only take 2 values (Bernoulli distribution), one value with probability p and the other with probability q = 1−p. For cases approaching this limit boundary with symmetry α = β, skewness ≈ 0, excess kurtosis ≈ −2 (this is the lowest excess kurtosis possible for any distribution), and the probabilities are $p \approx q \approx 1/2$. For cases approaching this limit boundary with skewness, excess kurtosis ≈ −2 + skewness^2, and the probability density is concentrated more at one end than the other end (with practically nothing in between), with probabilities $\boldsymbol{p = \frac{\beta}{\alpha+\beta}}$ at the left end x = 0 and $\boldsymbol{q = 1 - p = \frac{\alpha}{\alpha+\beta}}$ at the right end x = 1.

Symmetry

All statements are conditional on α, β > 0

- **Probability density function** reflection symmetry

$$f(x;\alpha,\beta) = f(1-x;\beta,\alpha)$$

- **Cumulative distribution function** reflection symmetry plus unitary translation

$$F(x;\alpha,\beta) = I_x(\alpha,\beta) = 1 - F(1-x;\beta,\alpha) = 1 - I_{1-x}(\beta,\alpha)$$

- **Mode** reflection symmetry plus unitary translation

$$\mathrm{mode}(\mathrm{B}(\alpha,\beta)) = 1 - \mathrm{mode}(\mathrm{B}(\beta,\alpha)),\ \text{if}\ \mathrm{B}(\beta,\alpha) \neq \mathrm{B}(1,1)$$

- **Median** reflection symmetry plus unitary translation

$$\text{median}(\text{B}(\alpha, \beta)) = 1 - \text{median}(\text{B}(\beta, \alpha))$$

- **Mean** reflection symmetry plus unitary translation

$$\mu(\text{B}(\alpha, \beta)) = 1 - \mu(\text{B}(\beta, \alpha))$$

- **Geometric Means** each is individually asymmetric, the following symmetry applies between the geometric mean based on X and the geometric mean based on its reflection (1-X)

$$G_X(\text{B}(\alpha, \beta)) = G_{(1-X)}(\text{B}(\beta, \alpha))$$

- **Harmonic means** each is individually asymmetric, the following symmetry applies between the harmonic mean based on X and the harmonic mean based on its reflection (1-X)

$$H_X(\text{B}(\alpha, \beta)) = H_{(1-X)}(\text{B}(\beta, \alpha)) \text{ if } \alpha, \beta > 1 \text{ .}$$

- **Variance** symmetry

$$\text{var}(\text{B}(\alpha, \beta)) = \text{var}(\text{B}(\beta, \alpha))$$

- **Geometric variances** each is individually asymmetric, the following symmetry applies between the log geometric variance based on X and the log geometric variance based on its reflection (1-X)

$$\ln(\text{var}_{\text{GX}}(\text{B}(\alpha, \beta))) = \ln(\text{var}_{\text{G(1-X)}}(\text{B}(\beta, \alpha)))$$

- **Geometric covariance** symmetry

$$\ln \text{ cov}_{\text{GX,(1-X)}}(\text{B}(\alpha, \beta)) = \ln \text{cov}_{\text{GX,(1-X)}}(\text{B}(\beta, \alpha))$$

- **Mean absolute deviation around the mean** symmetry

$$\text{E}[|X - E[X]|](\text{B}(\alpha, \beta)) = \text{E}[|X - E[X]|](\text{B}(\beta, \alpha))$$

- **Skewness** skew-symmetry

$$\text{skewness}(\text{B}(\alpha, \beta)) = -\text{skewness}(\text{B}(\beta, \alpha))$$

- **Excess kurtosis** symmetry

$$\text{excess kurtosis}(\mathrm{B}(\alpha,\beta)) = \text{excess kurtosis}(\mathrm{B}(\beta,\alpha))$$

- **Characteristic function** symmetry of Real part (with respect to the origin of variable "t")

$$\mathrm{Re}[{}_1F_1(\alpha;\alpha+\beta;it)] = \mathrm{Re}[{}_1F_1(\alpha;\alpha+\beta;-it)]$$

- **Characteristic function** skew-symmetry of Imaginary part (with respect to the origin of variable "t")

$$\mathrm{Im}[{}_1F_1(\alpha;\alpha+\beta;it)] = -\mathrm{Im}[{}_1F_1(\alpha;\alpha+\beta;-it)]$$

- **Characteristic function** symmetry of Absolute value (with respect to the origin of variable "t")

$$\mathrm{Abs}[{}_1F_1(\alpha;\alpha+\beta;it)] = \mathrm{Abs}[{}_1F_1(\alpha;\alpha+\beta;-it)]$$

- **Differential entropy** symmetry

$$h(\mathrm{B}(\alpha,\beta)) = h(\mathrm{B}(\beta,\alpha))$$

- **Relative Entropy (also called Kullback–Leibler divergence)** symmetry

$$D_{\mathrm{KL}}(X_1, X_2) = D_{\mathrm{KL}}(X_2, X_1), \text{ if } h(X_1) = h(X_2), \text{ for (skewed) } \alpha \neq \beta$$

- **Fisher information matrix** symmetry

$$\mathcal{I}_{i,j} = \mathcal{I}_{j,i}$$

Geometry of the probability density function

Inflection points

For certain values of the shape parameters α and β, the probability density function has inflection points, at which the curvature changes sign. The position of these inflection points can be useful as a measure of the dispersion or spread of the distribution.

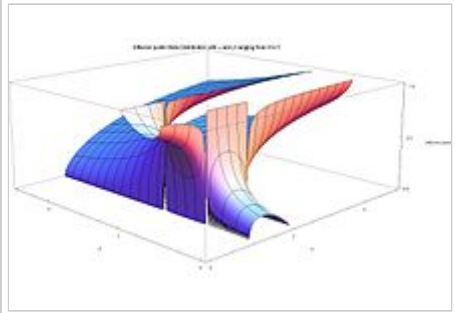
Defining the following quantity:

$$\kappa = \frac{\sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-3}}}{\alpha+\beta-2}$$

Points of inflection occur,[7][9][12][13] depending on the value of the shape parameters α and β, as follows:

- (α > 2, β > 2) The distribution is bell-shaped (symmetric for α = β and skewed otherwise), with **two inflection points**, equidistant from the mode:

$$x = \text{mode} \pm \kappa = \frac{\alpha - 1 \pm \sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-3}}}{\alpha + \beta - 2}$$



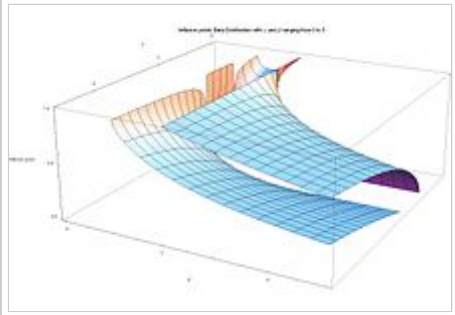
Inflection point location versus α and β showing regions with one inflection point

- (α = 2, β > 2) The distribution is unimodal, positively skewed, right-tailed, with **one inflection point**, located to the right of the mode:

$$x = \text{mode} + \kappa = \frac{2}{\beta}$$

- (α > 2, β = 2) The distribution is unimodal, negatively skewed, left-tailed, with **one inflection point**, located to the left of the mode:

$$x = \text{mode} - \kappa = 1 - \frac{2}{\alpha}$$



Inflection point location versus α and β showing region with two inflection points

- (1 < α < 2, β > 2, α+β>2) The distribution is unimodal, positively skewed, right-tailed, with **one inflection point**, located to the right of the mode:

$$x = \text{mode} + \kappa = \frac{\alpha - 1 + \sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-3}}}{\alpha + \beta - 2}$$

- (0 < α < 1, 1 < β < 2) The distribution has a mode at the left end $x = 0$ and it is positively skewed, right-tailed. There is **one inflection point**, located to the right of the mode:

$$x = \frac{\alpha - 1 + \sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-3}}}{\alpha + \beta - 2}$$

- (α > 2, 1 < β < 2) The distribution is unimodal negatively skewed, left-tailed, with **one inflection point**, located to the left of the mode:

$$x = \text{mode} - \kappa = \frac{\alpha - 1 - \sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-3}}}{\alpha + \beta - 2}$$

- (1 < α < 2, 0 < β < 1) The distribution has a mode at the right end x=1 and it is negatively skewed, left-tailed. There is **one inflection point**, located to the left of the mode:

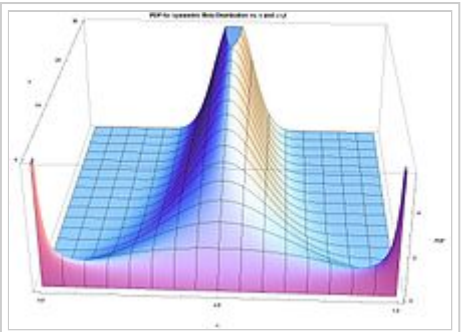
$$x = \frac{\alpha - 1 - \sqrt{\frac{(\alpha-1)(\beta-1)}{\alpha+\beta-3}}}{\alpha + \beta - 2}$$

There are no inflection points in the remaining (symmetric and skewed) regions: U-shaped: (α, β < 1) upside-down-U-shaped: (1 < α < 2, 1 < β < 2), reverse-J-shaped (α < 1, β > 2) or J-shaped: (α > 2, β < 1)

The accompanying plots show the inflection point locations (shown vertically, ranging from 0 to 1) versus α and β (the horizontal axes ranging from 0 to 5). There are large cuts at surfaces intersecting the lines α = 1, β = 1, α = 2, and β = 2 because at these values the beta distribution change from 2 modes, to 1 mode to no mode.

Shapes

The beta density function can take a wide variety of different shapes depending on the values of the two parameters α and β. The ability of the beta distribution to take this great diversity of shapes (using only two parameters) is partly responsible for finding wide application for modeling actual measurements:



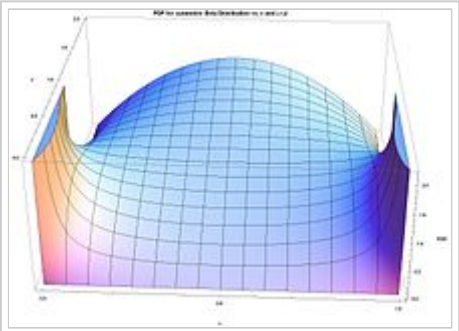
PDF for symmetric beta distribution vs. x and alpha=beta from 0 to 30

Symmetric (α = β)

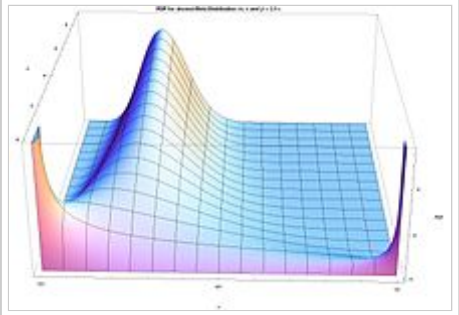
- the density function is symmetric about 1/2 (blue & teal plots).
- median = mean = 1/2.
- skewness = 0.
- **α = β < 1**
  - U-shaped (blue plot).
  - bimodal: left mode = 0, right mode =1, anti-mode = 1/2
  - 1/12 < var(X) < 1/4[7]
  - −2 < excess kurtosis(X) < −6/5
  - α = β = 1/2 is the arcsine distribution
    - var(X) = 1/8
    - excess kurtosis(X) = −3/2

- α = β → 0 is a 2-point Bernoulli distribution with equal probability 1/2 at each Dirac delta function end $x = 0$ and $x = 1$ and zero probability everywhere else. A coin toss: one face of the coin being $x = 0$ and the other face being $x = 1$.
  - $\lim_{\alpha=\beta\to 0} \operatorname{var}(X) = \frac{1}{4}$
  - $\lim_{\alpha=\beta\to 0} \text{excess kurtosis}(X) = -2$ a lower value than this is impossible for any distribution to reach.
  - The differential entropy approaches a minimum value of −∞

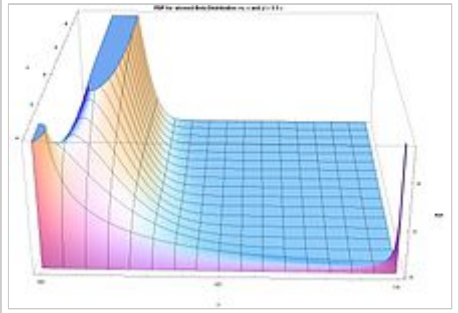
- **α = β = 1**
  - the uniform [0, 1] distribution
  - no mode
  - var(X) = 1/12
  - excess kurtosis(X) = −6/5
  - The (negative anywhere else) differential entropy reaches its maximum value of zero
- **α = β > 1**
  - symmetric unimodal
  - mode = 1/2.
  - 0 < var(X) < 1/12[7]
  - −6/5 < excess kurtosis(X) < 0
  - α = β = 3/2 is a semi-elliptic [0, 1] distribution, see: Wigner semicircle distribution
    - var(X) = 1/16.
    - excess kurtosis(X) = −1
  - α = β = 2 is the parabolic [0, 1] distribution
    - var(X) = 1/20
    - excess kurtosis(X) = −6/7
  - α = β > 2 is bell-shaped, with inflection points located to either side of the mode
    - 0 < var(X) < 1/20
    - −6/7 < excess kurtosis(X) < 0
  - α = β → ∞ is a 1-point Degenerate distribution with a Dirac delta function spike at the midpoint $x = 1/2$ with probability 1, and zero probability everywhere else. There is 100% probability (absolute certainty) concentrated at the single point $x = 1/2$.
    - $\lim_{\alpha=\beta\to\infty} \operatorname{var}(X) = 0$
    - $\lim_{\alpha=\beta\to\infty} \text{excess kurtosis}(X) = 0$
    - The differential entropy approaches a minimum value of −∞



PDF for symmetric beta distribution vs. x and alpha=beta from 0 to 2



PDF for skewed beta distribution vs. x and beta= 2.5 alpha from 0 to 9

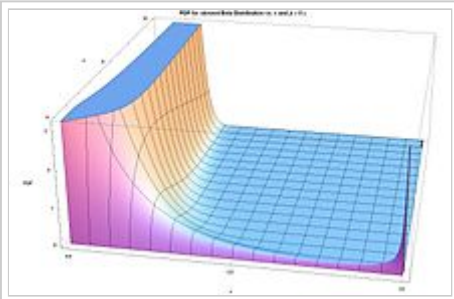


PDF for skewed beta distribution vs. x and beta= 5.5 alpha from 0 to 9

Skewed (α ≠ β)

The density function is skewed. An interchange of parameter values yields the mirror image (the reverse) of the initial curve, some more specific cases:

- **$\alpha < 1, \beta < 1$**
  - U-shaped
  - Positive skew for $\alpha < \beta$, negative skew for $\alpha > \beta$.
  - bimodal: left mode = 0, right mode = 1, anti-mode = $\frac{\alpha-1}{\alpha+\beta-2}$
  - $0 <$ median < 1.
  - $0 < \operatorname{var}(X) < 1/4$
- **$\alpha > 1, \beta > 1$**
  - unimodal (magenta & cyan plots),
  - Positive skew for $\alpha < \beta$, negative skew for $\alpha > \beta$.
  - $\textbf{mode} = \frac{\alpha-1}{\alpha+\beta-2}$
  - $0 <$ median < 1
  - $0 < \operatorname{var}(X) < 1/12$
- **$\alpha < 1, \beta \geq 1$**
  - reverse J-shaped with a right tail,
  - positively skewed,
  - strictly decreasing, convex
  - mode = 0
  - $0 <$ median $< 1/2$.
  - $\mathbf{0 < \operatorname{var}(X) < \frac{-11+5\sqrt{5}}{2}}$, (maximum variance occurs for $\alpha = \frac{-1+\sqrt{5}}{2}, \beta = 1$, or $\alpha = \Phi$ the golden ratio conjugate)
- **$\alpha \geq 1, \beta < 1$**
  - J-shaped with a left tail,
  - negatively skewed,
  - strictly increasing, convex
  - mode = 1
  - $1/2 <$ median < 1
  - $\mathbf{0 < \operatorname{var}(X) < \frac{-11+5\sqrt{5}}{2}}$, (maximum variance occurs for $\alpha = 1, \beta = \frac{-1+\sqrt{5}}{2}$, or $\beta = \Phi$ the golden ratio conjugate)
- **$\alpha = 1, \beta > 1$**
  - positively skewed,
  - strictly decreasing (red plot),
  - a reversed (mirror-image) power function [0,1] distribution
  - mode = 0
  - $\alpha = 1$, $1 < \beta < 2$
    - concave
    - $\mathbf{1 - \frac{1}{\sqrt{2}} < \text{median} < \frac{1}{2}}$
    - $1/18 < \operatorname{var}(X) < 1/12$.
  - $\alpha = 1$, $\beta = 2$



PDF for skewed beta distribution vs. x and beta= 8 alpha from 0 to 10

- a straight line with slope −2, the right-triangular distribution with right angle at the left end, at $x = 0$
- $\mathbf{median} = 1 - \frac{1}{\sqrt{2}}$
- var(X) = 1/18

- α = 1, β > 2
  - reverse J-shaped with a right tail,
  - convex
  - $\mathbf{0 < median} < 1 - \frac{1}{\sqrt{2}}$
  - 0 < var(X) < 1/18

- **α > 1, β = 1**
  - negatively skewed,
  - strictly increasing (green plot),
  - the power function [0, 1] distribution[12]
  - mode =1
  - 2 > α > 1, β = 1
    - concave
    - $\frac{1}{2} < \mathbf{median} < \frac{1}{\sqrt{2}}$
    - 1/18 < var(X) < 1/12
  - α = 2, β = 1
    - a straight line with slope +2, the right-triangular distribution with right angle at the right end, at $x = 1$
    - $\mathbf{median} = \frac{1}{\sqrt{2}}$
    - var(X) = 1/18
  - α > 2, β = 1
    - J-shaped with a left tail, convex
    - $\frac{1}{\sqrt{2}} < \mathbf{median} < 1$
    - 0 < var(X) < 1/18

Parameter estimation

Method of moments

Two unknown parameters

Two unknown parameters ($(\hat{\alpha}, \hat{\beta})$ of a beta distribution supported in the [0,1] interval) can be estimated, using the method of moments, with the first two moments (sample mean and sample variance) as follows. Let:

$$\text{sample mean}(\mathrm{X}) = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

be the sample mean estimate and

$$\text{sample variance}(\mathrm{X}) = \bar{v} = \frac{1}{N-1}\sum_{i=1}^{N} (X_i - \bar{x})^2$$

be the sample variance estimate. The method-of-moments estimates of the parameters are

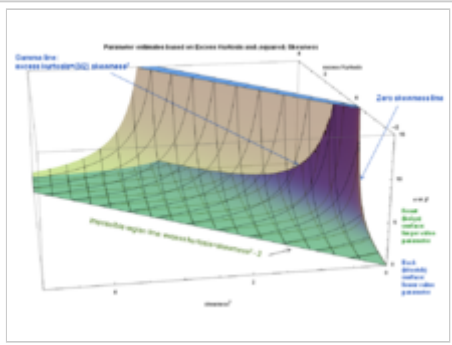
$$\hat{\alpha} = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right), \text{ if } \bar{v} < \bar{x}(1-\bar{x}),$$
$$\hat{\beta} = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right), \text{ if } \bar{v} < \bar{x}(1-\bar{x}).$$

When the distribution is required over known interval other than [0, 1] with random variable *X*, say [*a*, *c*] with random variable *Y*, then replace $\bar{x}$ with $\frac{\bar{y}-a}{c-a}$, and $\bar{v}$ with $\frac{v_{\bar{Y}}}{(c-a)^2}$ in the above couple of equations for the shape parameters (see "Alternative parametrizations, four parameters" section below).,[39] where:

$$\text{sample mean}(\mathrm{Y}) = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$
$$\text{sample variance}(\mathrm{Y}) = v_{\bar{Y}} = \frac{1}{N-1}\sum_{i=1}^{N} (Y_i - \bar{y})^2$$

Four unknown parameters

All four parameters ($\hat{\alpha}, \hat{\beta}, \hat{a}, \hat{c}$ of a beta distribution supported in the [*a*, *c*] interval -see section "Alternative parametrizations, Four parameters"-) can be estimated, using the method of moments developed by Karl Pearson, by equating sample and population values of the first four central moments (mean, variance, skewness and excess kurtosis).[7][40][41] The excess kurtosis was expressed in terms of the square of the skewness, and the sample size ν = α + β, (see previous section "Kurtosis") as follows:



Solutions for parameter estimates vs. (sample) excess Kurtosis and (sample) squared Skewness Beta distribution

$$\text{excess kurtosis} = \frac{6}{3+\nu}\left(\frac{(2+\nu)}{4}(\text{skewness})^2 - 1\right) \text{ if } (\text{skewness})^2 - 2 < \text{excess kurtosis} < \tfrac{3}{2}(\text{skewness})^2$$

One can use this equation to solve for the sample size ν= α + β in terms of the square of the skewness and the excess kurtosis as follows:[40]

$$\hat{\nu} = \hat{\alpha} + \hat{\beta} = 3\frac{(\text{sample excess kurtosis}) - (\text{sample skewness})^2 + 2}{\frac{3}{2}(\text{sample skewness})^2 - (\text{sample excess kurtosis})}$$
$$\text{if } (\text{sample skewness})^2 - 2 < \text{sample excess kurtosis} < \tfrac{3}{2}(\text{sample skewness})^2$$

This is the ratio (multiplied by a factor of 3) between the previously derived limit boundaries for the beta distribution in a space (as originally done by Karl Pearson[27]) defined with coordinates of the square of the skewness in one axis and the excess kurtosis in the other axis (see previous section titled "Kurtosis bounded by the square of the skewness"):

The case of zero skewness, can be immediately solved because for zero skewness, α = β and hence ν = 2α = 2β, therefore α = β = ν/2

$$\hat{\alpha} = \hat{\beta} = \frac{\hat{\nu}}{2} = \frac{\frac{3}{2}(\text{sample excess kurtosis}) + 3}{-(\text{sample excess kurtosis})}$$
$$\text{if sample skewness} = 0 \text{ and } -2 < \text{sample excess kurtosis} < 0$$

(Excess kurtosis is negative for the beta distribution with zero skewness, ranging from -2 to 0, so that $\hat{\nu}$ -and therefore the sample shape parameters- is positive, ranging from zero when the shape parameters approach zero and the excess kurtosis approaches -2, to infinity when the shape parameters approach infinity and the excess kurtosis approaches zero).

For non-zero sample skewness one needs to solve a system of two coupled equations. Since the skewness and the excess kurtosis are independent of the parameters $\hat{\alpha}, \hat{c}$, the parameters $\hat{\alpha}, \hat{\beta}$ can be uniquely determined from the sample skewness and the sample excess kurtosis, by solving the coupled equations with two known variables (sample skewness and sample excess kurtosis) and two unknowns (the shape parameters):

$$(\text{sample skewness})^2 = \frac{4(\hat{\beta}-\hat{\alpha})^2(1+\hat{\alpha}+\hat{\beta})}{\hat{\alpha}\hat{\beta}(2+\hat{\alpha}+\hat{\beta})^2}$$

$$\text{sample excess kurtosis} = \frac{6}{3+\hat{\alpha}+\hat{\beta}}\left(\frac{(2+\hat{\alpha}+\hat{\beta})}{4}(\text{sample skewness})^2 - 1\right)$$

$$\text{if } (\text{sample skewness})^2 - 2 < \text{sample excess kurtosis} < \tfrac{3}{2}(\text{sample skewness})^2$$

resulting in the following solution:[40]

$$\hat{\alpha}, \hat{\beta} = \frac{\hat{\nu}}{2}\left(1 \pm \frac{1}{\sqrt{1+\frac{16(\hat{\nu}+1)}{(\hat{\nu}+2)^2(\text{sample skewness})^2}}}\right)$$

$$\text{if sample skewness} \neq 0 \text{ and } (\text{sample skewness})^2 - 2 < \text{sample excess kurtosis} < \tfrac{3}{2}(\text{sample skewness})^2$$

Where one should take the solutions as follows: $\hat{\alpha} > \hat{\beta}$ for (negative) sample skewness < 0, and $\hat{\alpha} < \hat{\beta}$ for (positive) sample skewness > 0.

The accompanying plot shows these two solutions as surfaces in a space with horizontal axes of (sample excess kurtosis) and (sample squared skewness) and the shape parameters as the vertical axis. The surfaces are constrained by the condition that the sample excess kurtosis must be bounded by the sample squared skewness as stipulated in the above equation. The two surfaces meet at the right edge defined by zero skewness. Along this right edge, both parameters are equal and the distribution is symmetric U-shaped for α = β < 1, uniform for α = β = 1, upside-down-U-shaped for 1 < α = β < 2 and bell-shaped for α = β > 2. The surfaces also meet at the front (lower) edge defined by "the impossible boundary" line (excess kurtosis + 2 - skewness^2 = 0). Along this front (lower) boundary both shape parameters approach zero, and the probability density is concentrated more at one end than the other end (with practically nothing in between), with probabilities $p = \frac{\beta}{\alpha+\beta}$ at the left end $x = 0$ and $q = 1 - p = \frac{\alpha}{\alpha+\beta}$ at the right end $x = 1$. The two surfaces become further apart towards the rear edge. At this rear edge the surface parameters are quite different from each other. As remarked, for example, by Bowman and Shenton,[42] sampling in the neighborhood of the line (sample excess kurtosis - (3/2)(sample skewness)^2 = 0) (the just-J-shaped portion of the rear edge where blue meets beige), "is dangerously near to chaos", because at that line the denominator of the expression above for the estimate ν = α + β becomes zero and hence ν approaches infinity as that line is approached. Bowman and Shenton [42] write that "the higher moment parameters (kurtosis and skewness) are extremely fragile (near that line). However the mean and standard deviation are fairly reliable." Therefore, the problem is for the case of four parameter estimation for very skewed distributions such that the excess kurtosis approaches (3/2)

times the square of the skewness. This boundary line is produced by extremely skewed distributions with very large values of one of the parameters and very small values of the other parameter. See section titled "Kurtosis bounded by the square of the skewness" for a numerical example and further comments about this rear edge boundary line (sample excess kurtosis - (3/2)(sample skewness)^2 = 0). As remarked by Karl Pearson himself [43] this issue may not be of much practical importance as this trouble arises only for very skewed J-shaped (or mirror-image J-shaped) distributions with very different values of shape parameters that are unlikely to occur much in practice). The usual skewed skewed-bell-shape distributions that occur in practice do not have this parameter estimation problem.

The remaining two parameters $\hat{a}, \hat{c}$ can be determined using the sample mean and the sample variance using a variety of equations.[7][40] One alternative is to calculate the support interval range $(\hat{c} - \hat{a})$ based on the sample variance and the sample kurtosis. For this purpose one can solve, in terms of the range $(\hat{c} - \hat{a})$, the equation expressing the excess kurtosis in terms of the sample variance, and the sample size ν (see section titled "Kurtosis" and "Alternative parametrizations, four parameters"):

$$\text{sample excess kurtosis} = \frac{6}{(3+\hat{\nu})(2+\hat{\nu})}\left(\frac{(\hat{c}-\hat{a})^2}{(\text{sample variance})} - 6 - 5\hat{\nu}\right)$$

to obtain:

$$(\hat{c}-\hat{a}) = \sqrt{(\text{sample variance})}\sqrt{6+5\hat{\nu}+\frac{(2+\hat{\nu})(3+\hat{\nu})}{6}(\text{sample excess kurtosis})}$$

Another alternative is to calculate the support interval range $(\hat{c} - \hat{a})$ based on the sample variance and the sample skewness.[40] For this purpose one can solve, in terms of the range $(\hat{c} - \hat{a})$, the equation expressing the squared skewness in terms of the sample variance, and the sample size ν (see section titled "Skewness" and "Alternative parametrizations, four parameters"):

$$(\text{sample skewness})^2 = \frac{4}{(2+\hat{\nu})^2}\left(\frac{(\hat{c}-\hat{a})^2}{(\text{sample variance})} - 4(1+\hat{\nu})\right)$$

to obtain:[40]

$$(\hat{c}-\hat{a}) = \frac{\sqrt{(\text{sample variance})}}{2}\sqrt{(2+\hat{\nu})^2(\text{sample skewness})^2 + 16(1+\hat{\nu})}$$

The remaining parameter can be determined from the sample mean and the previously obtained parameters: $(\hat{c} - \hat{a}), \hat{\alpha}, \hat{\nu} = \hat{\alpha} + \hat{\beta}$:

$$\hat{a} = (\text{sample mean}) - \left(\frac{\hat{\alpha}}{\hat{\nu}}\right)(\hat{c}-\hat{a})$$

and finally, of course, $\hat{c} = (\hat{c} - \hat{a}) + \hat{a}$.

In the above formulas one may take, for example, as estimates of the sample moments:

$$\text{sample mean} = \overline{y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$$\text{sample variance} = \overline{v}_Y = \frac{1}{N-1}\sum_{i=1}^{N} (Y_i - \overline{y})^2$$

$$\text{sample skewness} = G_1 = \frac{N}{(N-1)(N-2)} \frac{\sum_{i=1}^{N} (Y_i - \overline{y})^3}{\overline{v}_Y^{\frac{3}{2}}}$$

$$\text{sample excess kurtosis} = G_2 = \frac{N(N+1)}{(N-1)(N-2)(N-3)} \frac{\sum_{i=1}^{N} (Y_i - \overline{y})^4}{\overline{v}_Y^2} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

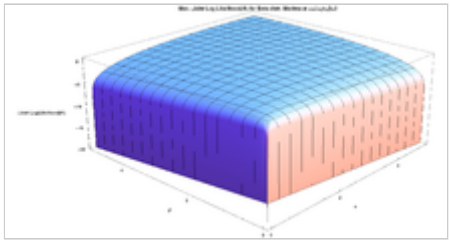
The estimators G_1 for sample skewness and G_2 for sample kurtosis are used by DAP/SAS, PSPP/SPSS, and Excel. However, they are not used by BMDP and (according to [44]) they were not used by MINITAB in 1998. Actually, Joanes and Gill in their 1998 study[44] concluded that the skewness and kurtosis estimators used in BMDP and in MINITAB (at that time) had smaller variance and mean-squared error in normal samples, but the skewness and kurtosis estimators used in DAP/SAS, PSPP/SPSS, namely G_1 and G_2, had smaller mean-squared error in samples from a very skewed distribution. It is for this reason that we have spelled out "sample skewness", etc., in the above formulas, to make it explicit that the user should choose the best estimator according to the problem at hand, as the best estimator for skewness and kurtosis depends on the amount of skewness (as shown by Joanes and Gill[44]).

Maximum likelihood

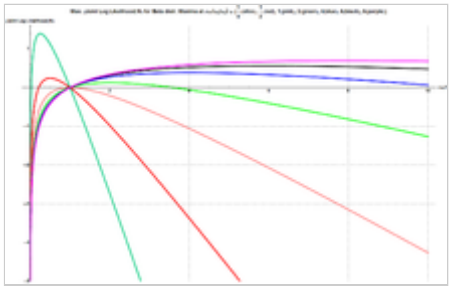
Two unknown parameters

As it is also the case for maximum likelihood estimates for the gamma distribution, the maximum likelihood estimates for the beta distribution do not have a general closed form solution for arbitrary values of the shape parameters. If X_1, ..., X_N are independent random variables each having a beta distribution, the joint log likelihood function for N iid observations is:

$$\begin{aligned}\ln \mathcal{L}(\alpha,\beta|X) &= \sum_{i=1}^{N} \ln(\mathcal{L}_i(\alpha,\beta|X_i)) \\ &= \sum_{i=1}^{N} \ln(f(X_i;\alpha,\beta)) \\ &= \sum_{i=1}^{N} \ln\left(\frac{X_i^{\alpha-1}(1-X_i)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\right) \\ &= (\alpha-1)\sum_{i=1}^{N}\ln(X_i) + (\beta-1)\sum_{i=1}^{N}\ln(1-X_i) - N\ln\mathrm{B}(\alpha,\beta)\end{aligned}$$



Max (Joint Log Likelihood/N) for Beta distribution Maxima at alpha=beta=2



Max (Joint Log Likelihood/N) for Beta distribution Maxima at alpha=beta= {0.25,0.5,1,2,4,6,8}

Finding the maximum with respect to a shape parameter involves taking the partial derivative with respect to the shape parameter and setting the expression equal to zero yielding the maximum likelihood estimator of the shape parameters:

$$\begin{aligned}\frac{\partial \ln \mathcal{L}(\alpha,\beta|X)}{\partial\alpha} &= \sum_{i=1}^{N}\ln X_i - N\frac{\partial \ln \mathrm{B}(\alpha,\beta)}{\partial\alpha} = 0 \\ \frac{\partial \ln \mathcal{L}(\alpha,\beta|X)}{\partial\beta} &= \sum_{i=1}^{N}\ln(1-X_i) - N\frac{\partial \ln \mathrm{B}(\alpha,\beta)}{\partial\beta} = 0\end{aligned}$$

where:

$$\begin{aligned}\frac{\partial \ln \mathrm{B}(\alpha,\beta)}{\partial\alpha} &= -\frac{\partial \ln \Gamma(\alpha+\beta)}{\partial\alpha} + \frac{\partial \ln \Gamma(\alpha)}{\partial\alpha} + \frac{\partial \ln \Gamma(\beta)}{\partial\alpha} = -\psi(\alpha+\beta) + \psi(\alpha) + 0 \\ \frac{\partial \ln \mathrm{B}(\alpha,\beta)}{\partial\beta} &= -\frac{\partial \ln \Gamma(\alpha+\beta)}{\partial\beta} + \frac{\partial \ln \Gamma(\alpha)}{\partial\beta} + \frac{\partial \ln \Gamma(\beta)}{\partial\beta} = -\psi(\alpha+\beta) + 0 + \psi(\beta)\end{aligned}$$

since the **digamma function** denoted ψ(α) is defined as the logarithmic derivative of the gamma function:[25]

$$\psi(\alpha) = \frac{\partial \ln \Gamma(\alpha)}{\partial\alpha}$$

To ensure that the values with zero tangent slope are indeed a maximum (instead of a saddle-point or a minimum) one has to also satisfy the condition that the curvature is negative. This amounts to satisfying that the second partial derivative with respect to the shape parameters is negative

$$\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta|X)}{\partial\alpha^2} = -N\frac{\partial^2 \ln \mathrm{B}(\alpha,\beta)}{\partial\alpha^2} < 0$$
$$\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta|X)}{\partial\beta^2} = -N\frac{\partial^2 \ln \mathrm{B}(\alpha,\beta)}{\partial\beta^2} < 0$$

using the previous equations, this is equivalent to:

$$\frac{\partial^2 \ln \mathrm{B}(\alpha,\beta)}{\partial\alpha^2} = \psi_1(\alpha) - \psi_1(\alpha+\beta) > 0$$
$$\frac{\partial^2 \ln \mathrm{B}(\alpha,\beta)}{\partial\beta^2} = \psi_1(\beta) - \psi_1(\alpha+\beta) > 0$$

where the **trigamma function**, denoted $\psi_1(\alpha)$, is the second of the polygamma functions, and is defined as the derivative of the digamma function:

$$\psi_1(\alpha) = \frac{\partial^2 \ln \Gamma(\alpha)}{\partial\alpha^2} = \frac{\partial \psi(\alpha)}{\partial\alpha}.$$

These conditions are equivalent to stating that the variances of the logarithmically transformed variables are positive, since:

$$\operatorname{var}[\ln(X)] = \mathrm{E}[\ln^2(X)] - (\mathrm{E}[\ln(X)])^2 = \psi_1(\alpha) - \psi_1(\alpha+\beta)$$
$$\operatorname{var}[\ln(1-X)] = \mathrm{E}[\ln^2(1-X)] - (\mathrm{E}[\ln(1-X)])^2 = \psi_1(\beta) - \psi_1(\alpha+\beta)$$

Therefore, the condition of negative curvature at a maximum is equivalent to the statements:

$$\operatorname{var}[\ln(X)] > 0$$
$$\operatorname{var}[\ln(1-X)] > 0$$

Alternatively, the condition of negative curvature at a maximum is also equivalent to stating that the following logarithmic derivatives of the geometric means G_X and $G_{(1-X)}$ are positive, since:

$$\psi_1(\alpha) - \psi_1(\alpha+\beta) = \frac{\partial \ln G_X}{\partial\alpha} > 0$$
$$\psi_1(\beta) - \psi_1(\alpha+\beta) = \frac{\partial \ln G_{(1-X)}}{\partial\beta} > 0$$

While these slopes are indeed positive, the other slopes are negative:

$$\frac{\partial \ln G_X}{\partial \beta}, \frac{\partial \ln G_{(1-X)}}{\partial \alpha} < 0.$$

The slopes of the mean and the median with respect to α and β display similar sign behavior.

From the condition that at a maximum, the partial derivative with respect to the shape parameter equals zero, we obtain the following system of coupled maximum likelihood estimate equations (for the average log-likelihoods) that needs to be inverted to obtain the (unknown) shape parameter estimates $\hat{\alpha}, \hat{\beta}$ in terms of the (known) average of logarithms of the samples X_1, ..., X_N:[7]

$$\hat{\mathbf{E}}[\ln(X)] = \psi(\hat{\alpha}) - \psi(\hat{\alpha} + \hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \ln X_i = \ln \hat{G}_X$$

$$\hat{\mathbf{E}}[\ln(1-X)] = \psi(\hat{\beta}) - \psi(\hat{\alpha} + \hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \ln(1-X_i) = \ln \hat{G}_{(1-X)}$$

where we recognize $\log \hat{G}_X$ as the logarithm of the sample geometric mean and $\log \hat{G}_{(1-X)}$ as the logarithm of the sample geometric mean based on (1-X), the mirror-image of X. For $\hat{\alpha} = \hat{\beta}$, it follows that $\hat{G}_X = \hat{G}_{(1-X)}$.

$$\hat{G}_X = \prod_{i=1}^{N} (X_i)^{\frac{1}{N}}$$

$$\hat{G}_{(1-X)} = \prod_{i=1}^{N} (1-X_i)^{\frac{1}{N}}$$

These coupled equations containing digamma functions of the shape parameter estimates $\hat{\alpha}, \hat{\beta}$ must be solved by numerical methods as done, for example, by Beckman et al.[45] Gnanadesikan et al. give numerical solutions for a few cases.[46] N.L.Johnson and S.Kotz[7] suggest that for "not too small" shape parameter estimates $\hat{\alpha}, \hat{\beta}$, the logarithmic approximation to the digamma function $\psi(\hat{\alpha}) \approx \ln(\hat{\alpha} - \frac{1}{2})$ may be used to obtain initial values for an iterative solution, since the equations resulting from this approximation can be solved exactly:

$$\ln \frac{\hat{\alpha} - \frac{1}{2}}{\hat{\alpha} + \hat{\beta} - \frac{1}{2}} \approx \ln \hat{G}_X$$

$$\ln \frac{\hat{\beta} - \frac{1}{2}}{\hat{\alpha} + \hat{\beta} - \frac{1}{2}} \approx \ln \hat{G}_{(1-X)}$$

which leads to the following solution for the initial values (of the estimate shape parameters in terms of the sample geometric means) for an iterative solution:

$$\hat{\alpha} \approx \tfrac{1}{2} + \frac{\hat{G}_X}{2(1 - \hat{G}_X - \hat{G}_{(1-X)})} \text{ if } \hat{\alpha} > 1$$
$$\hat{\beta} \approx \tfrac{1}{2} + \frac{\hat{G}_{(1-X)}}{2(1 - \hat{G}_X - \hat{G}_{(1-X)})} \text{ if } \hat{\beta} > 1$$

Alternatively, the estimates provided by the method of moments can instead be used as initial values for an iterative solution of the maximum likelihood coupled equations in terms of the digamma functions.

When the distribution is required over a known interval other than [0, 1] with random variable X, say [a, c] with random variable Y, then replace $\ln(X_i)$ in the first equation with

$$\ln \frac{Y_i - a}{c - a},$$

and replace $\ln(1-X_i)$ in the second equation with

$$\ln \frac{c - Y_i}{c - a}$$

(see "Alternative parametrizations, four parameters" section below).

If one of the shape parameters is known, the problem is considerably simplified. The following logit transformation can be used to solve for the unknown shape parameter (for skewed cases such that $\hat{\alpha} \neq \hat{\beta}$, otherwise, if symmetric, both -equal- parameters are known when one is known):

$$\hat{\mathrm{E}}\left[\ln\left(\frac{X}{1-X}\right)\right] = \psi(\hat{\alpha}) - \psi(\hat{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \ln \frac{X_i}{1-X_i} = \ln \hat{G}_X - \ln\left(\hat{G}_{(1-X)}\right)$$

This logit transformation is the logarithm of the transformation that divides the variable X by its mirror-image ($X/(1 - X)$ resulting in the "inverted beta distribution" or beta prime distribution (also known as beta distribution of the second kind or Pearson's Type VI) with support [0, +∞). As previously discussed in the section "Moments of logarithmically transformed random variables," the logit transformation $\ln \frac{X}{1-X}$, studied by Johnson,[31] extends the finite support [0, 1] based on the original variable X to infinite support in both directions of the real line (−∞, +∞).

If, for example, $\hat{\beta}$ is known, the unknown parameter $\hat{\alpha}$ can be obtained in terms of the inverse[47] digamma function of the right hand side of this equation:

$$\psi(\hat{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} \ln \frac{X_i}{1-X_i} + \psi(\hat{\beta})$$

$$\hat{\alpha} = (\text{Inverse digamma})[\ln \hat{G}_X - \ln \hat{G}_{(1-X)} + \psi(\hat{\beta})]$$

In particular, if one of the shape parameters has a value of unity, for example for $\hat{\beta} = 1$ (the power function distribution with bounded support [0,1]), using the identity ψ(x + 1) = ψ(x) + 1/x in the equation $\psi(\hat{\alpha}) - \psi(\hat{\alpha} + \hat{\beta}) = \ln \hat{G}_X$, the maximum likelihood estimator for the unknown parameter $\hat{\alpha}$ is,[7] exactly:

$$\hat{\alpha} = -\frac{1}{\frac{1}{N}\sum_{i=1}^{N} \ln X_i} = -\frac{1}{\ln \hat{G}_X}$$

The beta has support [0, 1], therefore $\hat{G}_X < 1$, and hence $(-\ln \hat{G}_X) > 0$, and therefore $\hat{\alpha} > 0$.

In conclusion, the maximum likelihood estimates of the shape parameters of a beta distribution are (in general) a complicated function of the sample geometric mean, and of the sample geometric mean based on *(1−X)*, the mirror-image of *X*. One may ask, if the variance (in addition to the mean) is necessary to estimate two shape parameters with the method of moments, why is the (logarithmic or geometric) variance not necessary to estimate two shape parameters with the maximum likelihood method, for which only the geometric means suffice? The answer is because the mean does not provide as much information as the geometric mean. For a beta distribution with equal shape parameters α = β, the mean is exactly 1/2, regardless of the value of the shape parameters, and therefore regardless of the value of the statistical dispersion (the variance). On the other hand, the geometric mean of a beta distribution with equal shape parameters α = β, depends on the value of the shape parameters, and therefore it contains more information. Also, the geometric mean of a beta distribution does not satisfy the symmetry conditions satisfied by the mean, therefore, by employing both the geometric mean based on *X* and geometric mean based on *(1−X)*, the maximum likelihood method is able to provide best estimates for both parameters α = β, without need of employing the variance.

One can express the joint log likelihood per *N* iid observations in terms of the *sufficient statistics* (the sample geometric means) as follows:

$$\frac{\ln \mathcal{L}(\alpha, \beta|X)}{N} = (\alpha - 1)\ln \hat{G}_X + (\beta - 1)\ln \hat{G}_{(1-X)} - \ln \mathrm{B}(\alpha, \beta)$$

We can plot the joint log likelihood per N observations for fixed values of the sample geometric means to see the behavior of the likelihood function as a function of the shape parameters α and β. In such a plot, the shape parameter estimators $\hat{\alpha}, \hat{\beta}$ correspond to the maxima of the likelihood function. See the accompanying graph that shows that all the likelihood functions intersect at α = β = 1, which corresponds to the values of the shape parameters that give the maximum entropy (the maximum entropy occurs for shape parameters equal to unity: the uniform distribution). It is evident from the plot that the likelihood function gives sharp peaks for values of the shape parameter estimators close to zero, but that for values of the shape parameters estimators greater than one, the likelihood function becomes quite flat, with less defined peaks. Obviously, the maximum likelihood parameter estimation method for the beta distribution becomes less acceptable for larger values of the shape parameter estimators, as the uncertainty in the peak definition increases with the value of the shape parameter estimators. One can arrive at the same conclusion by noticing that the expression for the curvature of the likelihood function is in terms of the geometric variances

$$\frac{\partial^2 \ln \mathcal{L}(\alpha, \beta|X)}{\partial \alpha^2} = -\operatorname{var}[\ln X]$$
$$\frac{\partial^2 \ln \mathcal{L}(\alpha, \beta|X)}{\partial \beta^2} = -\operatorname{var}[\ln(1-X)]$$

These variances (and therefore the curvatures) are much larger for small values of the shape parameter α and β. However, for shape parameter values α, β > 1, the variances (and therefore the curvatures) flatten out. Equivalently, this result follows from the Cramér–Rao bound, since the Fisher information matrix components for the beta distribution are these logarithmic variances. The Cramér–Rao bound states that the variance of any *unbiased* estimator $\hat{\alpha}$ of α is bounded by the reciprocal of the Fisher information:

$$\operatorname{var}(\hat{\alpha}) \geq \frac{1}{\operatorname{var}[\ln X]} \geq \frac{1}{\psi_1(\hat{\alpha}) - \psi_1(\hat{\alpha} + \hat{\beta})}$$
$$\operatorname{var}(\hat{\beta}) \geq \frac{1}{\operatorname{var}[\ln(1-X)]} \geq \frac{1}{\psi_1(\hat{\beta}) - \psi_1(\hat{\alpha} + \hat{\beta})}$$

so the variance of the estimators increases with increasing α and β, as the logarithmic variances decrease.

Also one can express the joint log likelihood per N iid observations in terms of the digamma function expressions for the logarithms of the sample geometric means as follows:

$$\frac{\ln \mathcal{L}(\alpha, \beta|X)}{N} = (\alpha - 1)(\psi(\hat{\alpha}) - \psi(\hat{\alpha} + \hat{\beta})) + (\beta - 1)(\psi(\hat{\beta}) - \psi(\hat{\alpha} + \hat{\beta})) - \ln \mathrm{B}(\alpha, \beta)$$

this expression is identical to the negative of the cross-entropy (see section on "Quantities of information (entropy)"). Therefore, finding the maximum of the joint log likelihood of the shape parameters, per N iid observations, is identical to finding the minimum of the cross-entropy for the beta distribution, as a function of the shape parameters.

$$\frac{\ln\,\mathcal{L}(\alpha,\beta|X)}{N} = -H = -h - D_{\mathrm{KL}} = -\ln \mathrm{B}(\alpha,\beta) + (\alpha-1)\psi(\hat{\alpha}) + (\beta-1)\psi(\hat{\beta}) - (\alpha+\beta-2)\psi(\hat{\alpha}+\hat{\beta})$$

with the cross-entropy defined as follows:

$$H = \int_0^1 -f(X;\hat{\alpha},\hat{\beta})\ln(f(X;\alpha,\beta))\,\mathrm{d}X$$

Four unknown parameters

The procedure is similar to the one followed in the two unknown parameter case. If Y_1, ..., Y_N are independent random variables each having a beta distribution with four parameters, the joint log likelihood function for N iid observations is:

$$\begin{aligned}
\ln\,\mathcal{L}(\alpha,\beta,a,c|Y) &= \sum_{i=1}^{N} \ln\,\mathcal{L}_i(\alpha,\beta,a,c|Y_i) \\
&= \sum_{i=1}^{N} \ln\,f(Y_i;\alpha,\beta,a,c) \\
&= \sum_{i=1}^{N} \ln\,\frac{(Y_i-a)^{\alpha-1}(c-Y_i)^{\beta-1}}{(c-a)^{\alpha+\beta-1}\mathrm{B}(\alpha,\beta)} \\
&= (\alpha-1)\sum_{i=1}^{N}\ln(Y_i-a) + (\beta-1)\sum_{i=1}^{N}\ln(c-Y_i) - N\ln\mathrm{B}(\alpha,\beta) - N(\alpha+\beta-1)\ln(c-a)
\end{aligned}$$

Finding the maximum with respect to a shape parameter involves taking the partial derivative with respect to the shape parameter and setting the expression equal to zero yielding the maximum likelihood estimator of the shape parameters:

$$\begin{aligned}
\frac{\partial \ln\mathcal{L}(\alpha,\beta,a,c|Y)}{\partial\alpha} &= \sum_{i=1}^{N}\ln(Y_i-a) - N(-\psi(\alpha+\beta)+\psi(\alpha)) - N\ln(c-a) = 0 \\
\frac{\partial \ln\mathcal{L}(\alpha,\beta,a,c|Y)}{\partial\beta} &= \sum_{i=1}^{N}\ln(c-Y_i) - N(-\psi(\alpha+\beta)+\psi(\beta)) - N\ln(c-a) = 0 \\
\frac{\partial \ln\mathcal{L}(\alpha,\beta,a,c|Y)}{\partial a} &= -(\alpha-1)\sum_{i=1}^{N}\frac{1}{Y_i-a} + N(\alpha+\beta-1)\frac{1}{c-a} = 0
\end{aligned}$$

$$\frac{\partial \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial c} = (\beta-1)\sum_{i=1}^{N}\frac{1}{c-Y_i} - N(\alpha+\beta-1)\frac{1}{c-a} = 0$$

these equations can be re-arranged as the following system of four coupled equations (the first two equations are geometric means and the second two equations are the harmonic means) in terms of the maximum likelihood estimates for the four parameters $\hat{\alpha}, \hat{\beta}, \hat{a}, \hat{c}$:

$$\frac{1}{N}\sum_{i=1}^{N}\ln\frac{Y_i-\hat{a}}{\hat{c}-\hat{a}} = \psi(\hat{\alpha}) - \psi(\hat{\alpha}+\hat{\beta}) = \ln \hat{G}_X$$

$$\frac{1}{N}\sum_{i=1}^{N}\ln\frac{\hat{c}-Y_i}{\hat{c}-\hat{a}} = \psi(\hat{\beta}) - \psi(\hat{\alpha}+\hat{\beta}) = \ln \hat{G}_{1-X}$$

$$\frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{c}-\hat{a}}{Y_i-\hat{a}}} = \frac{\hat{\alpha}-1}{\hat{\alpha}+\hat{\beta}-1} = \hat{H}_X$$

$$\frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{c}-\hat{a}}{\hat{c}-Y_i}} = \frac{\hat{\beta}-1}{\hat{\alpha}+\hat{\beta}-1} = \hat{H}_{1-X}$$

with sample geometric means:

$$\hat{G}_X = \prod_{i=1}^{N}\left(\frac{Y_i-\hat{a}}{\hat{c}-\hat{a}}\right)^{\frac{1}{N}}$$

$$\hat{G}_{(1-X)} = \prod_{i=1}^{N}\left(\frac{\hat{c}-Y_i}{\hat{c}-\hat{a}}\right)^{\frac{1}{N}}$$

The parameters $\hat{a}, \hat{c}$ are embedded inside the geometric mean expressions in a nonlinear way (to the power $1/N$). This precludes, in general, a closed form solution, even for an initial value approximation for iteration purposes. One alternative is to use as initial values for iteration the values obtained from the method of moments solution for the four parameter case. Furthermore, the expressions for the harmonic means are well-defined only for $\hat{\alpha}, \hat{\beta} > 1$, which precludes a maximum likelihood solution for shape parameters less than unity in the four-parameter case. Fisher's information matrix for the four parameter case is positive-definite only for α, β > 2 (for further discussion, see section on Fisher information matrix, four parameter case), for bell-shaped (symmetric or unsymmetric) beta distributions, with inflection points located to either side of the mode. The following Fisher information components (that represent the expectations of the curvature of the log likelihood function) have singularities at the following values:

$$\alpha = 2: \quad \mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial a^2}\right] = \mathcal{I}_{a,a}$$

$$\beta = 2: \quad \mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial c^2}\right] = \mathcal{I}_{c,c}$$

$$\alpha = 2: \quad \mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \alpha \partial a}\right] = \mathcal{I}_{\alpha,a}$$

$$\beta = 1: \quad \mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \beta \partial c}\right] = \mathcal{I}_{\beta,c}$$

(for further discussion see section on Fisher information matrix). Thus, it is not possible to strictly carry on the maximum likelihood estimation for some well known distributions belonging to the four-parameter beta distribution family, like the uniform distribution (Beta(1, 1, *a*, *c*)), and the arcsine distribution (Beta(1/2, 1/2, *a*, *c*)). N.L.Johnson and S.Kotz[7] ignore the equations for the harmonic means and instead suggest "If a and c are unknown, and maximum likelihood estimators of *a*, *c*, α and β are required, the above procedure (for the two unknown parameter case, with *X* transformed as $X = (Y-a)/(c-a)$) can be repeated using a succession of trial values of *a* and *c*, until the pair (*a*, *c*) for which maximum likelihood (given *a* and *c*) is as great as possible, is attained" (where, for the purpose of clarity, their notation for the parameters has been translated into the present notation).

Fisher information matrix

Let a random variable X have a probability density f(x;α). The partial derivative with respect to the (unknown, and to be estimated) parameter α of the log likelihood function is called the score. The second moment of the score is called the Fisher information:

$$\mathcal{I}(\alpha) = \mathbf{E}\left[\left(\frac{\partial}{\partial \alpha}\ln \mathcal{L}(\alpha|X)\right)^2\right],$$

The expectation of the score is zero, therefore the Fisher information is also the second moment centered on the mean of the score: the variance of the score.

If the log likelihood function is twice differentiable with respect to the parameter α, and under certain regularity conditions,[48] then the Fisher information may also be written as follows (which is often a more convenient form for calculation purposes):

$$\mathcal{I}(\alpha) = -\mathbf{E}\left[\frac{\partial^2}{\partial \alpha^2}\ln(\mathcal{L}(\alpha|X))\right].$$

Thus, the Fisher information is the negative of the expectation of the second derivative with respect to the parameter α of the log likelihood function. Therefore, Fisher information is a measure of the curvature of the log likelihood function of α. A low curvature (and therefore high radius of curvature), flatter log likelihood function curve has low Fisher information; while a log likelihood function curve with large curvature (and therefore low radius of curvature) has high Fisher information. When the Fisher information matrix is computed at the evaluates of the parameters ("the observed Fisher information matrix") it is equivalent to the replacement of the true log likelihood surface by a Taylor's series approximation, taken as far as the quadratic terms.[49] The word information, in the context of Fisher information, refers to information about the parameters. Information such as: estimation, sufficiency and properties of variances of estimators. The Cramér–Rao bound states that the inverse of the Fisher information is a lower bound on the variance of any estimator of a parameter α:

$$\mathbf{var}[\hat{\alpha}] \geq \frac{1}{\mathcal{I}(\alpha)}.$$

The precision to which one can estimate the estimator of a parameter α is limited by the Fisher Information of the log likelihood function. The Fisher information is a measure of the minimum error involved in estimating a parameter of a distribution and it can be viewed as a measure of the resolving power of an experiment needed to discriminate between two alternative hypothesis of a parameter.[50]

When there are *N* parameters

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_N \end{bmatrix},$$

then the Fisher information takes the form of an *N*×*N* positive semidefinite symmetric matrix, the Fisher Information Matrix, with typical element:

$$(\mathcal{I}(\theta))_{i,j} = \mathbf{E}\left[\left(\frac{\partial}{\partial\theta_i}\ln\mathcal{L}\right)\left(\frac{\partial}{\partial\theta_j}\ln\mathcal{L}\right)\right].$$

Under certain regularity conditions,[48] the Fisher Information Matrix may also be written in the following form, which is often more convenient for computation:

$$(\mathcal{I}(\theta))_{i,j} = -\mathbf{E}\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ln(\mathcal{L})\right].$$

With X_1, ..., X_N iid random variables, an *N*-dimensional "box" can be constructed with sides X_1, ..., X_N. Costa and Cover[51] show that the (Shannon) differential entropy $h(X)$ is related to the volume of the typical set (having the sample entropy close to the true entropy), while the Fisher information is related to the surface of this typical set.

Two parameters

For X_1, ..., X_N independent random variables each having a beta distribution parametrized with shape parameters α and β, the joint log likelihood function for N iid observations is:

$$\ln(\mathcal{L}(\alpha,\beta|X)) = (\alpha-1)\sum_{i=1}^{N}\ln X_i + (\beta-1)\sum_{i=1}^{N}\ln(1-X_i) - N\ln \mathrm{B}(\alpha,\beta)$$

therefore the joint log likelihood function per N iid observations is:

$$\frac{1}{N}\ln(\mathcal{L}(\alpha,\beta|X)) = (\alpha-1)\frac{1}{N}\sum_{i=1}^{N}\ln X_i + (\beta-1)\frac{1}{N}\sum_{i=1}^{N}\ln(1-X_i) - \ln \mathrm{B}(\alpha,\beta)$$

For the two parameter case, the Fisher information has 4 components: 2 diagonal and 2 off-diagonal. Since the Fisher information matrix is symmetric, one of these off diagonal components is independent. Therefore, the Fisher information matrix has 3 independent components (2 diagonal and 1 off diagonal).

Aryal and Nadarajah[52] calculated Fisher's information matrix for the four parameter case, from which the two parameter case can be obtained as follows:

$$-\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta|X)}{N\partial\alpha^2} = \operatorname{var}[\ln(X)] = \psi_1(\alpha) - \psi_1(\alpha+\beta) = \mathcal{I}_{\alpha,\alpha} = \mathrm{E}\left[-\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta|X)}{N\partial\alpha^2}\right] = \ln \operatorname{var}_{\mathrm{GX}}$$

$$-\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta|X)}{N\partial\beta^2} = \operatorname{var}[\ln(1-X)] = \psi_1(\beta) - \psi_1(\alpha+\beta) = \mathcal{I}_{\beta,\beta} = \mathrm{E}\left[-\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta|X)}{N\partial\beta^2}\right] = \ln \operatorname{var}_{\mathrm{G(1-X)}}$$

$$-\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta|X)}{N\partial\alpha\partial\beta} = \operatorname{cov}[\ln X, \ln(1-X)] = -\psi_1(\alpha+\beta) = \mathcal{I}_{\alpha,\beta} = \mathrm{E}\left[-\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta|X)}{N\partial\alpha\partial\beta}\right] = \ln \operatorname{cov}_{\mathrm{GX,(1-X)}}$$

Since the Fisher information matrix is symmetric

$$\mathcal{I}_{\alpha,\beta} = \mathcal{I}_{\beta,\alpha} = \ln \operatorname{cov}_{\mathrm{GX,(1-X)}}$$

The Fisher information components are equal to the log geometric variances and log geometric covariance. Therefore, they can be expressed as **trigamma functions**, denoted $\psi_1(\alpha)$, the second of the polygamma functions, defined as the derivative of the digamma function:

$$\psi_1(\alpha) = \frac{d^2 \ln\Gamma(\alpha)}{\partial\alpha^2} = \frac{\partial\psi(\alpha)}{\partial\alpha}.$$

These derivatives are also derived in the section titled "Parameter estimation", "Maximum likelihood", "Two unknown parameters," and plots of the log likelihood function are also shown in that section. The section titled "Geometric variance and covariance" contains plots and further discussion of the Fisher information matrix components: the log geometric variances and log geometric covariance as a function of the shape parameters α and β. The section titled "Other moments", "Moments of transformed random variables", "Moments of logarithmically transformed random variables" contains formulas for moments of logarithmically transformed random variables. Images for the Fisher information components $\mathcal{I}_{\alpha,\alpha}$, $\mathcal{I}_{\beta,\beta}$ and $\mathcal{I}_{\alpha,\beta}$ are shown in the section titled "Geometric variance".

The determinant of Fisher's information matrix is of interest (for example for the calculation of Jeffreys prior probability). From the expressions for the individual components of the Fisher information matrix, it follows that the determinant of Fisher's (symmetric) information matrix for the beta distribution is:

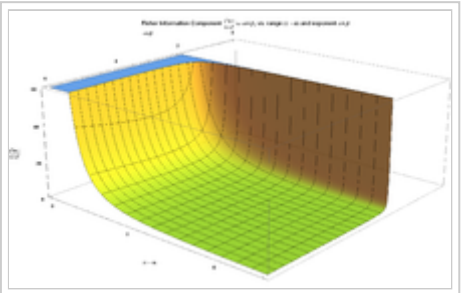
$$\begin{aligned}
\det(\mathcal{I}(\alpha,\beta)) &= \mathcal{I}_{\alpha,\alpha}\mathcal{I}_{\beta,\beta} - \mathcal{I}_{\alpha,\beta}\mathcal{I}_{\alpha,\beta} \\
&= (\psi_1(\alpha) - \psi_1(\alpha+\beta))(\psi_1(\beta) - \psi_1(\alpha+\beta)) - (-\psi_1(\alpha+\beta))(-\psi_1(\alpha+\beta)) \\
&= \psi_1(\alpha)\psi_1(\beta) - (\psi_1(\alpha) + \psi_1(\beta))\psi_1(\alpha+\beta) \\
\lim_{\alpha\to 0} \det(\mathcal{I}(\alpha,\beta)) &= \lim_{\beta\to 0} \det(\mathcal{I}(\alpha,\beta)) = \infty \\
\lim_{\alpha\to \infty} \det(\mathcal{I}(\alpha,\beta)) &= \lim_{\beta\to \infty} \det(\mathcal{I}(\alpha,\beta)) = 0
\end{aligned}$$

From Sylvester's criterion (checking whether the diagonal elements are all positive), it follows that the Fisher information matrix for the two parameter case is positive-definite (under the standard condition that the shape parameters are positive α>0 and β>0).

Four parameters

If Y_1, ..., Y_N are independent random variables each having a beta distribution with four parameters: the exponents α and β, as well as "a" (the minimum of the distribution range), and "c" (the maximum of the distribution range) (section titled "Alternative parametrizations", "Four parameters"), with probability density function:

$$f(y;\alpha,\beta,a,c) = \frac{f(x;\alpha,\beta)}{c-a} = \frac{\left(\frac{y-a}{c-a}\right)^{\alpha-1}\left(\frac{c-y}{c-a}\right)^{\beta-1}}{(c-a)B(\alpha,\beta)} = \frac{(y-a)^{\alpha-1}(c-y)^{\beta-1}}{(c-a)^{\alpha+\beta-1}B(\alpha,\beta)}.$$

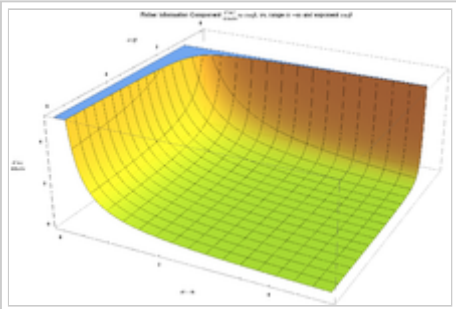


Fisher Information I(a,a) for α=β vs range (c-a) and exponent α=β

the joint log likelihood function per N iid observations is:

$$\frac{1}{N}\ln(\mathcal{L}(\alpha,\beta,a,c|Y)) = \frac{\alpha-1}{N}\sum_{i=1}^{N}\ln(Y_i-a) + \frac{\beta-1}{N}\sum_{i=1}^{N}\ln(c-Y_i) - \ln \mathrm{B}(\alpha,\beta) - (\alpha+\beta-1)\ln(c-a)$$

For the four parameter case, the Fisher information has 4*4=16 components. It has 12 off-diagonal components = (4*4 total - 4 diagonal). Since the Fisher information matrix is symmetric, half of these components (12/2=6) are independent. Therefore, the Fisher information matrix has 6 independent off-diagonal + 4 diagonal = 10 independent components. Aryal and Nadarajah[52] calculated Fisher's information matrix for the four parameter case as follows:



Fisher Information I(α,a) for α=β, vs. range (c - a) and exponent α=β

$$-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \alpha^2} = \operatorname{var}[\ln(X)] = \psi_1(\alpha) - \psi_1(\alpha+\beta) = \mathcal{I}_{\alpha,\alpha} = \mathrm{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \alpha^2}\right] = \ln(\operatorname{var}_{\mathrm{GX}})$$

$$-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \beta^2} = \operatorname{var}[\ln(1-X)] = \psi_1(\beta) - \psi_1(\alpha+\beta) = \mathcal{I}_{\beta,\beta} = \mathrm{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \beta^2}\right] = \ln(\operatorname{var}_{\mathrm{G(1-X)}})$$

$$-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \alpha \partial \beta} = \operatorname{cov}[\ln X, (1-X)] = -\psi_1(\alpha+\beta) = \mathcal{I}_{\alpha,\beta} = \mathrm{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \alpha \partial \beta}\right] = \ln(\operatorname{cov}_{\mathrm{GX,(1-X)}})$$

In the above expressions, the use of X instead of Y in the expressions $\operatorname{var}[\ln(X)] = \ln(\operatorname{var}_{\mathrm{GX}})$ is *not an error*. The expressions in terms of the log geometric variances and log geometric covariance occur as functions of the two parameter $X \sim \text{Beta}(\alpha, \beta)$ parametrization because when taking the partial derivatives with respect to the exponents (α, β) in the four parameter case, one obtains the identical expressions as for the two parameter case: these terms of the four parameter Fisher information matrix are independent of the minimum "a" and maximum "c" of the distribution's range. The only non-zero term upon double differentiation of the log likelihood function with respect to the exponents α and β is the second derivative of the log of the beta function: ln(B(α, β)). This term is independent of the minimum "a" and maximum "c" of the distribution's range. Double differentiation of this term results in trigamma functions. The sections titled "Maximum likelihood", "Two unknown parameters" and "Four unknown parameters" also show this fact.

The Fisher information for N i.i.d. samples is N times the individual Fisher information (eq. 11.279, page 394 of Cover and Thomas[34]). (Aryal and Nadarajah[52] take a single observation, $N = 1$, to calculate the following components of the Fisher information, which leads to the same result as considering the derivatives of the log likelihood per N observations. Moreover, below the erroneous expression for $\mathcal{I}_{a,a}$ in Aryal and Nadarajah has been corrected.)

$$\alpha > 2: \quad \mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial a^2}\right] = \mathcal{I}_{a,a} = \frac{\beta(\alpha+\beta-1)}{(\alpha-2)(c-a)^2}$$

$$\beta > 2: \quad \mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial c^2}\right] = \mathcal{I}_{c,c} = \frac{\alpha(\alpha+\beta-1)}{(\beta-2)(c-a)^2}$$

$$\mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial a \partial c}\right] = \mathcal{I}_{a,c} = \frac{(\alpha+\beta-1)}{(c-a)^2}$$

$$\alpha > 1: \quad \mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \alpha \partial a}\right] = \mathcal{I}_{\alpha,a} = \frac{\beta}{(\alpha-1)(c-a)}$$

$$\mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \alpha \partial c}\right] = \mathcal{I}_{\alpha,c} = \frac{1}{(c-a)}$$

$$\mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \beta \partial a}\right] = \mathcal{I}_{\beta,a} = -\frac{1}{(c-a)}$$

$$\beta > 1: \quad \mathbf{E}\left[-\frac{1}{N}\frac{\partial^2 \ln \mathcal{L}(\alpha,\beta,a,c|Y)}{\partial \beta \partial c}\right] = \mathcal{I}_{\beta,c} = -\frac{\alpha}{(\beta-1)(c-a)}$$

The lower two diagonal entries of the Fisher information matrix, with respect to the parameter "a" (the minimum of the distribution's range): $\mathcal{I}_{a,a}$, and with respect to the parameter "c" (the maximum of the distribution's range): $\mathcal{I}_{c,c}$ are only defined for exponents α > 2 and β > 2 respectively. The Fisher information matrix component $\mathcal{I}_{a,a}$ for the minimum "a" approaches infinity for exponent α approaching 2 from above, and the Fisher information matrix component $\mathcal{I}_{c,c}$ for the maximum "c" approaches infinity for exponent β approaching 2 from above.

The Fisher information matrix for the four parameter case does not depend on the individual values of the minimum "a" and the maximum "c", but only on the total range (*c*−*a*). Moreover, the components of the Fisher information matrix that depend on the range (*c*−*a*), depend only through its inverse (or the square of the inverse), such that the Fisher information decreases for increasing range (*c*−*a*).

The accompanying images show the Fisher information components $\mathcal{I}_{a,a}$ and $\mathcal{I}_{\alpha,a}$. Images for the Fisher information components $\mathcal{I}_{\alpha,\alpha}$ and $\mathcal{I}_{\beta,\beta}$ are shown in the section titled "Geometric variance". All these Fisher information components look like a basin, with the "walls" of the basin being located at low values of the parameters.

The following four-parameter-beta-distribution Fisher information components can be expressed in terms of the two-parameter: $X \sim$ Beta(α, β) expectations of the transformed ratio ((1-*X*)/*X*) and of its mirror image (*X*/(1-*X*)), scaled by the range (*c*−*a*), which may be helpful for interpretation:

$$\mathcal{I}_{\alpha,a} = \frac{\mathbf{E}\left[\frac{1-X}{X}\right]}{c-a} = \frac{\beta}{(\alpha-1)(c-a)} \text{ if } \alpha > 1$$
$$\mathcal{I}_{\beta,c} = -\frac{\mathbf{E}\left[\frac{X}{1-X}\right]}{c-a} = -\frac{\alpha}{(\beta-1)(c-a)} \text{ if } \beta > 1$$

These are also the expected values of the "inverted beta distribution" or beta prime distribution (also known as beta distribution of the second kind or Pearson's Type VI) [7] and its mirror image, scaled by the range (*c*−*a*).

Also, the following Fisher information components can be expressed in terms of the harmonic (1/X) variances or of variances based on the ratio transformed variables ((1-X)/X) as follows:

$$\alpha > 2: \quad \mathcal{I}_{a,a} = \mathbf{var}\left[\frac{1}{X}\right]\left(\frac{\alpha-1}{c-a}\right)^2 = \mathbf{var}\left[\frac{1-X}{X}\right]\left(\frac{\alpha-1}{c-a}\right)^2 = \frac{\beta(\alpha+\beta-1)}{(\alpha-2)(c-a)^2}$$
$$\beta > 2: \quad \mathcal{I}_{c,c} = \mathbf{var}\left[\frac{1}{1-X}\right]\left(\frac{\beta-1}{c-a}\right)^2 = \mathbf{var}\left[\frac{X}{1-X}\right]\left(\frac{\beta-1}{c-a}\right)^2 = \frac{\alpha(\alpha+\beta-1)}{(\beta-2)(c-a)^2}$$
$$\mathcal{I}_{a,c} = \mathbf{cov}\left[\frac{1}{X}, \frac{1}{1-X}\right]\frac{(\alpha-1)(\beta-1)}{(c-a)^2} = \mathbf{cov}\left[\frac{1-X}{X}, \frac{X}{1-X}\right]\frac{(\alpha-1)(\beta-1)}{(c-a)^2} = \frac{(\alpha+\beta-1)}{(c-a)^2}$$

See section "Moments of linearly transformed, product and inverted random variables" for these expectations.

The determinant of Fisher's information matrix is of interest (for example for the calculation of Jeffreys prior probability). From the expressions for the individual components, it follows that the determinant of Fisher's (symmetric) information matrix for the beta distribution with four parameters is:

$$\begin{aligned}\det(\mathcal{I}(\alpha,\beta,a,c)) = {} & -\mathcal{I}_{a,c}^2\mathcal{I}_{\alpha,a}\mathcal{I}_{\alpha,\beta} + \mathcal{I}_{a,a}\mathcal{I}_{a,c}\mathcal{I}_{\alpha,c}\mathcal{I}_{\alpha,\beta} + \mathcal{I}_{a,c}^2\mathcal{I}_{\alpha,\beta}^2 - \mathcal{I}_{a,a}\mathcal{I}_{c,c}\mathcal{I}_{\alpha,\beta}^2 \\ & - \mathcal{I}_{a,c}\mathcal{I}_{\alpha,a}\mathcal{I}_{\alpha,c}\mathcal{I}_{\beta,a} + \mathcal{I}_{a,c}^2\mathcal{I}_{\alpha,\alpha}\mathcal{I}_{\beta,a} + 2\mathcal{I}_{c,c}\mathcal{I}_{\alpha,a}\mathcal{I}_{\alpha,\beta}\mathcal{I}_{\beta,a} \\ & - 2\mathcal{I}_{a,c}\mathcal{I}_{\alpha,c}\mathcal{I}_{\alpha,\beta}\mathcal{I}_{\beta,a} + \mathcal{I}_{\alpha,c}^2\mathcal{I}_{\beta,a}^2 - \mathcal{I}_{c,c}\mathcal{I}_{\alpha,\alpha}\mathcal{I}_{\beta,a}^2 + \mathcal{I}_{a,c}\mathcal{I}_{\alpha,a}^2\mathcal{I}_{\beta,c} \\ & - \mathcal{I}_{a,a}\mathcal{I}_{a,c}\mathcal{I}_{\alpha,\alpha}\mathcal{I}_{\beta,c} - \mathcal{I}_{a,c}\mathcal{I}_{\alpha,a}\mathcal{I}_{\alpha,\beta}\mathcal{I}_{\beta,c} + \mathcal{I}_{a,a}\mathcal{I}_{\alpha,c}\mathcal{I}_{\alpha,\beta}\mathcal{I}_{\beta,c} \\ & - \mathcal{I}_{\alpha,a}\mathcal{I}_{\alpha,c}\mathcal{I}_{\beta,a}\mathcal{I}_{\beta,c} + \mathcal{I}_{a,c}\mathcal{I}_{\alpha,\alpha}\mathcal{I}_{\beta,a}\mathcal{I}_{\beta,c} - \mathcal{I}_{c,c}\mathcal{I}_{\alpha,a}^2\mathcal{I}_{\beta,\beta} \\ & + 2\mathcal{I}_{a,c}\mathcal{I}_{\alpha,a}\mathcal{I}_{\alpha,c}\mathcal{I}_{\beta,\beta} - \mathcal{I}_{a,a}\mathcal{I}_{\alpha,c}^2\mathcal{I}_{\beta,\beta} - \mathcal{I}_{a,c}^2\mathcal{I}_{\alpha,\alpha}\mathcal{I}_{\beta,\beta} + \mathcal{I}_{a,a}\mathcal{I}_{c,c}\mathcal{I}_{\alpha,\alpha}\mathcal{I}_{\beta,\beta} \text{ if } \alpha,\beta > 2\end{aligned}$$

Using Sylvester's criterion (checking whether the diagonal elements are all positive), and since diagonal components $\mathcal{I}_{a,a}$ and $\mathcal{I}_{c,c}$ have singularities at α=2 and β=2 it follows that the Fisher information matrix for the four parameter case is positive-definite for α>2 and β>2. Since for α > 2 and β > 2 the beta distribution is (symmetric or unsymmetric) bell shaped, it follows that the Fisher information matrix is positive-definite only for bell-shaped (symmetric or unsymmetric) beta distributions, with inflection points located to either side of the mode. Thus, important well known distributions belonging to the four-parameter beta distribution family, like the parabolic distribution (Beta(2,2,a,c)) and the uniform distribution (Beta(1,1,a,c)) have Fisher information components ($\mathcal{I}_{a,a}, \mathcal{I}_{c,c}, \mathcal{I}_{\alpha,a}, \mathcal{I}_{\beta,c}$) that blow up (approach infinity) in the four-parameter case (although their Fisher information components are all defined for the two parameter case). The four-parameter Wigner semicircle distribution (Beta(3/2,3/2,a,c)) and arcsine distribution (Beta(1/2,1/2,a,c)) have negative Fisher information determinants for the four-parameter case.

Generating beta-distributed random variates

If X and Y are independent, with $X \sim \Gamma(\alpha, \theta)$ and $Y \sim \Gamma(\beta, \theta)$ then

$$\frac{X}{X+Y} \sim \mathrm{B}(\alpha, \beta).$$

So one algorithm for generating beta variates is to generate $X/(X + Y)$, where X is a gamma variate with parameters (α, 1) and Y is an independent gamma variate with parameters (β, 1).[53]

Also, the kth order statistic of n uniformly distributed variates is $\mathrm{B}(k, n + 1 - k)$, so an alternative if α and β are small integers is to generate α + β − 1 uniform variates and choose the α-th smallest.[54]

Another way to generate the Beta distribution is by Pólya urn model. According to this method (http://www.tc.umn.edu/~horte005/docs/Dirichletdistribution.pdf), one start with an "urn" with α "black" balls and β "white" balls and draw uniformly with replacement. Every trial an additional ball is added according to the color of the last ball which was drawn. Asymptotically, the proportion of black and white balls will be distributed according to the Beta distribution, where each repetition of the experiment will produce a different value.

Related distributions

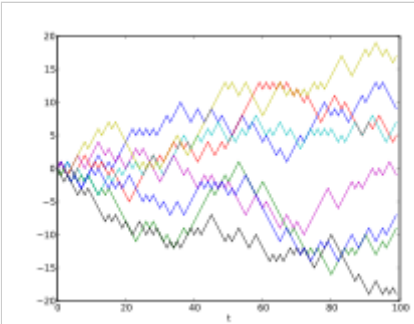
Transformations

- If $X \sim$ Beta(α, β) then $1-X \sim$ Beta(β, α) mirror-image symmetry
- If $X \sim$ Beta(α, β) then $\frac{X}{1-X} \sim \beta'(\alpha, \beta)$. The beta prime distribution, also called "beta distribution of the second kind".
- If $X \sim$ Beta(n/2, m/2) then $\frac{mX}{n(1-X)} \sim F(n, m)$ (assuming $n > 0$ and $m > 0$), the Fisher-Snedecor F distribution.

- If $X \sim Beta\left(1+\lambda\frac{m-\min}{\max-\min}, 1+\lambda\frac{\max-m}{\max-\min}\right)$ then min+X(max−min) ~ PERT(min, max, m, λ) where *PERT* denotes a distribution used in PERT analysis, and m=most likely value.[55] Traditionally[2] λ = 4 in PERT analysis.
- If X ~ Beta(1, β) then X ~ Kumaraswamy distribution with parameters (1, β)
- If X ~ Beta(α, 1) then X ~ Kumaraswamy distribution with parameters (α, 1)
- If X ~ Beta(α, 1) then −ln(X) ~ Exponential(α)

Special and limiting cases

- Beta(1, 1) ~ U(0, 1).
- If X ~ Beta(3/2, 3/2) and r > 0 then 2rX−r ~ Wigner semicircle distribution.
- Beta(1/2, 1/2) is equivalent to the arcsine distribution. This distribution is also Jeffreys prior probability for the Bernoulli and binomial distributions . The arcsine probability density is a distribution that appears in several random walk fundamental theorems. In a fair coin toss random walk, the probability for the time of the last visit to the origin is distributed as an (U-shaped) arcsine distribution.[11][19] In a two-player fair-coin-toss game, a player is said to be in the lead if the random walk (that started at the origin) is above the origin. The most probable number of times that a given player will be in the lead, in a game of length 2N, is not N. On the contrary, N is the least likely number of times that the player will be in the lead. The most likely number of times in the lead is 0 or 2N (following the arcsine distribution).
- $\lim_{n\to\infty} nBeta(1,n) = \mathbf{Exponential(1)}$ the exponential distribution
- $\lim_{n\to\infty} nBeta(k,n) = \mathbf{Gamma}(k,1)$ the gamma distribution



Example of eight realizations of a random walk in one dimension starting at 0: the probability for the time of the last visit to the origin is distributed as Beta(1/2, 1/2)

Derived from other distributions

- The kth order statistic of a sample of size n from the uniform distribution is a beta random variable, $U_{(k)}$ ~ Beta(k, n+1−k).[54]
- If X ~ Gamma(α, θ) and Y ~ Gamma(β, θ) are independent, then $\frac{X}{X+Y} \sim Beta(\alpha,\beta)$.
- If $X \sim \chi^2(\alpha)$ and $Y \sim \chi^2(\beta)$ are independent, then $\frac{X}{X+Y} \sim Beta(\frac{\alpha}{2},\frac{\beta}{2})$.
- If X ~ U(0, 1) and α > 0 then $X^{1/\alpha}$ ~ Beta(α, 1). The power function distribution.

Combination with other distributions

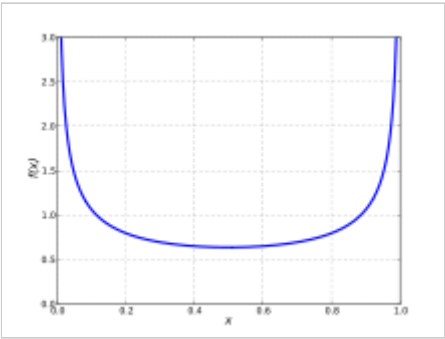
- X ~ Beta(α, β) and Y ~ F(2β,2α) then $\mathbf{Pr}(X \leq \frac{\alpha}{\alpha+\beta x}) = \mathbf{Pr}(Y \geq x)$ for all x > 0.

Compounding with other distributions

- If $p \sim \text{Beta}(\alpha, \beta)$ and $X \sim \text{Bin}(k, p)$ then $X \sim$ beta-binomial distribution
- If $p \sim \text{Beta}(\alpha, \beta)$ and $X \sim \text{NB}(r, p)$ then $X \sim$ beta negative binomial distribution

Generalisations

- The Dirichlet distribution is a multivariate generalization of the beta distribution. Univariate marginals of the Dirichlet distribution have a beta distribution. The beta distribution is conjugate to the binomial and Bernoulli distributions in exactly the same way as the Dirichlet distribution is conjugate to the multinomial distribution and categorical distribution.
- The Pearson type I distribution is identical to the beta distribution (except for arbitrary shifting and re-scaling that can also be accomplished with the four parameter parametrization of the beta distribution).
- $Beta(\alpha, \beta) = \lim_{\delta \to 0} \text{NonCentralBeta}(\alpha, \beta, \delta)$ the noncentral beta distribution
- the Generalized beta distribution is a five-parameter distribution family which has the beta distribution as a special case.



Beta(1/2, 1/2): The arcsine distribution probability density was proposed by Harold Jeffreys to represent uncertainty for a Bernoulli or a binomial distribution in Bayesian inference, and is now commonly referred to as Jeffreys prior: $p^{-1/2}(1-p)^{-1/2}$. This distribution also appears in several random walk fundamental theorems

Applications

Order statistics

The beta distribution has an important application in the theory of order statistics. A basic result is that the distribution of the kth smallest of a sample of size n from a continuous uniform distribution has a beta distribution.[54] This result is summarized as:

$$U_{(k)} \sim Beta(k, n+1-k).$$

From this, and application of the theory related to the probability integral transform, the distribution of any individual order statistic from any continuous distribution can be derived.[54]

Rule of succession

A classic application of the beta distribution is the rule of succession, introduced in the 18th century by Pierre-Simon Laplace[56] in the course of treating the sunrise problem. It states that, given s successes in n conditionally independent Bernoulli trials with probability p, that the estimate of the expected value in the next trial is $\frac{s+1}{n+2}$. This estimate is the expected value of the posterior distribution over p, namely Beta(s+1, n−s+1), which is given by Bayes' rule if one assumes a uniform prior probability over p (i.e., Beta(1, 1)) and then observes that p generated s successes in n trials. Laplace's rule of succession has been criticized by

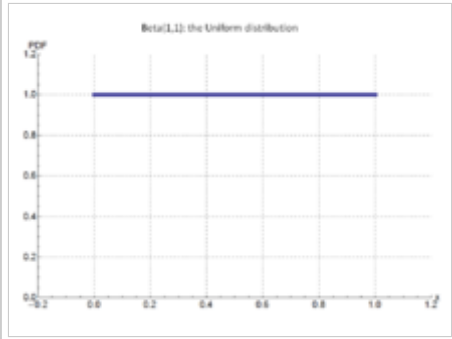
prominent scientists. R. T. Cox described Laplace's application of the rule of succession to the sunrise problem ([57] p. 89) as "a travesty of the proper use of the principle." Keynes remarks ([58] Ch.XXX, p. 382) "indeed this is so foolish a theorem that to entertain it is discreditable." Karl Pearson[59] showed that the probability that the next (n + 1) trials will be successes, after n successes in n trials, is only 50%, which has been considered too low by scientists like Jeffreys and unacceptable as a representation of the scientific process of experimentation to test a proposed scientific law. As pointed out by Jeffreys ([60] p. 128) (crediting C. D. Broad[61]) Laplace's rule of succession establishes a high probability of success ((n+1)/(n+2)) in the next trial, but only a moderate probability (50%) that a further sample (n+1) comparable in size will be equally successful. As pointed out by Perks,[62] "The rule of succession itself is hard to accept. It assigns a probability to the next trial which implies the assumption that the actual run observed is an average run and that we are always at the end of an average run. It would, one would think, be more reasonable to assume that we were in the middle of an average run. Clearly a higher value for both probabilities is necessary if they are to accord with reasonable belief." These problems with Laplace's rule of succession motivated Haldane, Perks, Jeffreys and others to search for other forms of prior probability (see the next section titled "Bayesian inference"). According to Jaynes,[50] the main problem with the rule of succession is that it is not valid when s=0 or s=n (see rule of succession, for an analysis of its validity).

Bayesian inference

The use of Beta distributions in Bayesian inference is due to the fact that they provide a family of conjugate prior probability distributions for binomial (including Bernoulli) and geometric distributions. The domain of the beta distribution can be viewed as a probability, and in fact the beta distribution is often used to describe the distribution of a probability value p:[30]

$$P(p;\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}.$$

Examples of beta distributions used as prior probabilities to represent ignorance of prior parameter values in Bayesian inference are Beta(1,1), Beta(0,0) and Beta(1/2,1/2).



Beta*(1, 1)**: The uniform distribution probability density was proposed by Thomas Bayes to represent ignorance of prior probabilities in Bayesian inference. It describes ***not a state of complete ignorance, but the state of knowledge in which we have observed at least one success and one failure, and therefore *we have prior knowledge that* ***both*** *states are physically possible.*

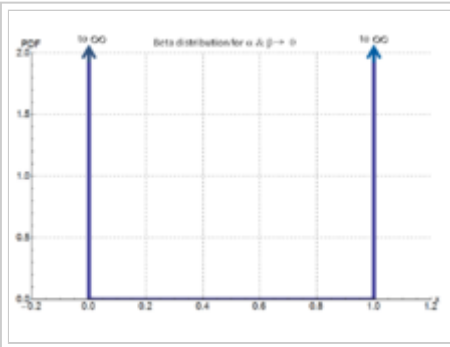
Bayes' prior probability (Beta(1,1))

The beta distribution achieves maximum differential entropy for Beta(1,1): the uniform probability density, for which all values in the domain of the distribution have equal density. This uniform distribution Beta(1,1) was suggested ("with a great deal of doubt") by Thomas Bayes[63] as the prior probability distribution to express ignorance about the correct prior distribution. This prior distribution was adopted (apparently, from his writings, with little sign of doubt[56]) by Pierre-Simon Laplace, and hence it was also known as the "Bayes-Laplace rule" or the "Laplace rule" of "inverse probability" in publications of the first half of the 20th century. In the later part of the 19th century and early part of the 20th century, scientists realized that the assumption of uniform "equal" probability density depended on the actual functions (for example whether a linear or a logarithmic scale was most appropriate) and parametrizations used. In particular, the behavior near the

ends of distributions with finite support (for example near $x = 0$, for a distribution with initial support at $x = 0$) required particular attention. Keynes ([58] Ch.XXX, p. 381) criticized the use of Bayes's uniform prior probability (Beta(1,1)) that all values between zero and one are equiprobable, as follows: "Thus experience, if it shows anything, shows that there is a very marked clustering of statistical ratios in the neighborhoods of zero and unity, of those for positive theories and for correlations between positive qualities in the neighborhood of zero, and of those for negative theories and for correlations between negative qualities in the neighborhood of unity. "

Haldane's prior probability (Beta(0,0))

The Beta(0,0) distribution was proposed by J.B.S. Haldane,[64] who suggested that the prior probability representing complete uncertainty should be proportional to $p^{-1}(1-p)^{-1}$. The function $p^{-1}(1-p)^{-1}$ can be viewed as the limit of the numerator of the beta distribution as both shape parameters approach zero: $\alpha, \beta \to 0$. The Beta function (in the denominator of the beta distribution) approaches infinity, for both parameters approaching zero, $\alpha, \beta \to 0$. Therefore, $p^{-1}(1-p)^{-1}$ divided by the Beta function approaches a 2-point Bernoulli distribution with equal probability 1/2 at each Dirac delta function end, at 0 and 1, and nothing in between, as $\alpha, \beta \to 0$. A coin-toss: one face of the coin being at 0 and the other face being at 1. The Haldane prior probability distribution Beta(0,0) is an "improper prior" because its integration (from 0 to 1) fails to strictly converge to 1 due to the Dirac delta function singularities at each end. However, this is not an issue for computing posterior probabilities unless the sample size is very small. Furthermore, Zellner[65] points out that on the log-odds scale, (the logit transformation $\ln(p/1-p)$), the Haldane prior is the uniformly flat prior. The fact that a uniform prior probability on the logit transformed variable $\ln(p/1-p)$ (with domain $(-\infty, \infty)$) is equivalent to the Haldane prior on the domain [0, 1] was pointed out by Harold Jeffreys in the first edition (1939) of his book Theory of Probability ([60] p. 123). Jeffreys writes "Certainly if we take the Bayes-Laplace rule right up to the extremes we are led to results that do not correspond to anybody's way of thinking. The (Haldane) rule $dx/(x(1-x))$ goes too far the other way. It would lead to the conclusion that if a sample is of one type with respect to some property there is a probability 1 that the whole population is of that type." The fact that "uniform" depends on the parametrization, led Jeffreys to seek a form of prior that would be invariant under different parametrizations.



***Beta*(0, 0)**: The Haldane prior probability expressing total ignorance about prior information, where we are not even sure whether it is physically possible for an experiment to yield either a success or a failure. As $\alpha, \beta \to 0$, the beta distribution approaches a two-point Bernoulli distribution with all probability density concentrated at each Dirac delta function end, at 0 and 1, and nothing in between. A coin-toss: one face of the coin being at 0 and the other face being at 1.

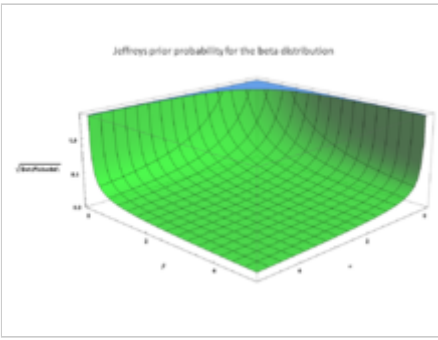
Jeffreys' prior probability (Beta(1/2,1/2) for a Bernoulli or for a binomial distribution)

Harold Jeffreys[60][66] proposed to use an uninformative prior probability measure that should be invariant under reparameterization: proportional to the square root of the determinant of Fisher's information matrix. For the Bernoulli distribution, this can be shown as follows: for a coin that is "heads" with probability $p \in [0, 1]$ and is "tails" with probability $1-p$, for a given (H,T) $\in \{(0,1), (1,0)\}$ the probability is $p^H(1-p)^T$. Since $T = 1-H$, the Bernoulli distribution is $p^H(1-p)^{1-H}$. Considering p as the only parameter, it follows that the log likelihood for the Bernoulli distribution is

$$\ln \mathcal{L}(p|H) = H \ln(p) + (1-H) \ln(1-p).$$

The Fisher information matrix has only one component (it is a scalar, because there is only one parameter: p), therefore:

$$\begin{aligned}\sqrt{\mathcal{I}(p)} &= \sqrt{\mathrm{E}\left[\left(\frac{d}{dp}\ln(\mathcal{L}(p|H))\right)^2\right]} \\ &= \sqrt{\mathrm{E}\left[\left(\frac{H}{p}-\frac{1-H}{1-p}\right)^2\right]} \\ &= \sqrt{p^1(1-p)^0\left(\frac{1}{p}-\frac{0}{1-p}\right)^2+p^0(1-p)^1\left(\frac{0}{p}-\frac{1}{1-p}\right)^2} \\ &= \frac{1}{\sqrt{p(1-p)}}.\end{aligned}$$



Jeffreys prior probability for the beta distribution: the square root of the determinant of Fisher's information matrix: $\sqrt{\det(\mathcal{I}(\alpha,\beta))}=\sqrt{\psi_1(\alpha)\psi_1(\beta)-(\psi_1(\alpha)+\psi_1(\beta))\psi_1}$ is a function of the trigamma function ψ_1 of shape parameters α, β

Similarly, for the Binomial distribution with n Bernoulli trials, it can be shown that

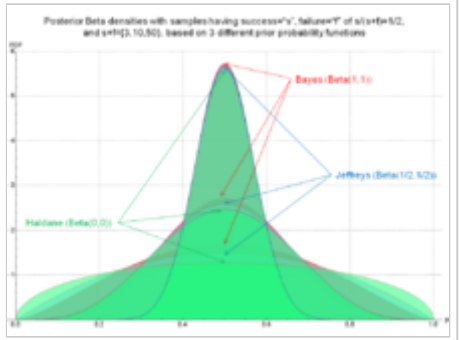
$$\sqrt{\mathcal{I}(p)} = \frac{\sqrt{n}}{\sqrt{p(1-p)}}.$$

Thus, for the Bernoulli, and Binomial distributions, Jeffreys prior is proportional to $\frac{1}{\sqrt{p(1-p)}}$, which happens to be proportional to a beta distribution with domain variable $x = p$, and shape parameters $\alpha = \beta = 1/2$, the arcsine distribution:

$$Beta(\tfrac{1}{2},\tfrac{1}{2}) = \frac{1}{\pi\sqrt{p(1-p)}}.$$

It will be shown in the next section that the normalizing constant for Jeffreys prior is immaterial to the final result because the normalizing constant cancels out in Bayes theorem for the posterior probability. Hence Beta(1/2,1/2) is used as the Jeffreys prior for both Bernoulli and binomial distributions. As shown in the next section, when using this expression as a prior probability times the likelihood in Bayes theorem, the posterior probability turns out to be a beta distribution. It is important to realize, however, that Jeffreys prior is proportional to $\frac{1}{\sqrt{p(1-p)}}$ for the Bernoulli and binomial distribution, but not for the beta distribution. Jeffreys prior for the beta distribution is given by the determinant of Fisher's information for the beta distribution, which, as shown in the section titled "Fisher information" is a function of the trigamma function ψ_1 of shape parameters α and β as follows:

$$\sqrt{\det(\mathcal{I}(\alpha,\beta))} = \sqrt{\psi_1(\alpha)\psi_1(\beta) - (\psi_1(\alpha)+\psi_1(\beta))\psi_1(\alpha+\beta)}$$
$$\lim_{\alpha\to 0}\sqrt{\det(\mathcal{I}(\alpha,\beta))} = \lim_{\beta\to 0}\sqrt{\det(\mathcal{I}(\alpha,\beta))} = \infty$$
$$\lim_{\alpha\to \infty}\sqrt{\det(\mathcal{I}(\alpha,\beta))} = \lim_{\beta\to \infty}\sqrt{\det(\mathcal{I}(\alpha,\beta))} = 0$$



Posterior Beta densities with samples having success="s", failure="f" of $s/(s+f)=1/2$, and $s+f=\{3,10,50\}$, based on 3 different prior probability functions: Haldane (Beta(0,0), Jeffreys (Beta(1/2,1/2)) and Bayes (Beta(1,1)). The image shows that there is little difference between the priors for the posterior with sample size of 50 (with more pronounced peak near $p=1/2$). Significant differences appear for very small sample sizes (the flatter distribution for sample size of 3)

As previously discussed, Jeffreys prior for the Bernoulli and binomial distributions is proportional to the arcsine distribution Beta(1/2,1/2), a one-dimensional *curve* that looks like a basin as a function of the parameter *p* of the Bernoulli and binomial distributions. The walls of the basin are formed by *p* approaching the singularities at the ends $p \to 0$ and $p \to 1$, where Beta(1/2,1/2) approaches infinity. Jeffreys prior for the beta distribution is a *2-dimensional surface* (embedded in a three-dimensional space) that looks like a basin with only two of its walls meeting at the corner $\alpha = \beta = 0$ (and missing the other two walls) as a function of the shape parameters α and β of the beta distribution. The two adjoining walls of this 2-dimensional surface are formed by the shape parameters α and β approaching the singularities (of the trigamma function) at $\alpha, \beta \to 0$. It has no walls for $\alpha, \beta \to \infty$ because in this case the determinant of Fisher's information matrix for the beta distribution approaches zero.

It will be shown in the next section that Jeffreys prior probability results in posterior probabilities (when multiplied by the binomial likelihood function) that are intermediate between the posterior probability results of the Haldane and Bayes prior probabilities.

Jeffreys prior may be difficult to obtain analytically, and for some cases it just doesn't exist (even for simple distribution functions like the asymmetric triangular distribution). Berger, Bernardo and Sun, in a 2009 paper[67] defined a reference prior probability distribution that (unlike Jeffreys prior) exists for the asymmetric triangular distribution. They cannot obtain a closed-form expression for their reference prior, but numerical calculations show it to be nearly perfectly fitted by the (proper) prior

$$Beta(\tfrac{1}{2},\tfrac{1}{2}) \sim \frac{1}{\sqrt{\theta(1-\theta)}}$$

where θ is the vertex variable for the asymmetric triangular distribution with support [0, 1] (corresponding to the following parameter values in Wikipedia's article on the triangular distribution: vertex c=θ, left end a=0,and right end b=1). Berger et al. also give a heuristic argument that Beta(1/2,1/2) could indeed be the exact Berger-Bernardo-Sun reference prior for the asymmetric triangular distribution. Therefore, Beta(1/2,1/2) not only is Jeffreys prior for the Bernoulli and binomial distributions, but also seems to be the Berger-Bernardo-Sun reference prior for the asymmetric triangular distribution (for which the Jeffreys prior does not exist), a distribution used in project management and PERT analysis to describe the cost and duration of project tasks.

Clarke and Barron[68] prove that, among continuous positive priors, Jeffreys prior (when it exists) asymptotically maximizes Shannon's mutual information between a sample of size n and the parameter, and therefore *Jeffreys prior is the most uninformative prior* (measuring information as Shannon information). The proof rests on an examination of the Kullback-Leibler distance between probability density functions for iid random variables.

Effect of different prior probability choices on the posterior beta distribution

If samples are drawn from the population of a random variable *X* that result in *s* successes and *f* failures in "n" Bernoulli trials n=*s*+*f*, then the likelihood function for parameters *s* and *f* given *x* = *p* (the notation *x*=*p* in the expressions below will emphasize that the domain *x* stands for the value of the parameter *p* in the binomial distribution), is the following binomial distribution:

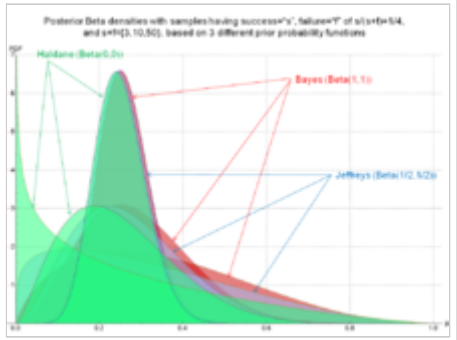
$$\mathcal{L}(s,f|x=p) = \binom{s+f}{s} x^s(1-x)^f = \binom{n}{s} x^s(1-x)^{n-s}.$$

If beliefs about prior probability information are reasonably well approximated by a beta distribution with parameters αPrior and βPrior, then:

$$\mathbf{PriorProbability}(x=p;\alpha\mathbf{Prior},\beta\mathbf{Prior}) = \frac{x^{\alpha\mathbf{Prior}-1}(1-x)^{\beta\mathbf{Prior}-1}}{\mathrm{B}(\alpha\mathbf{Prior},\beta\mathbf{Prior})}$$

According to Bayes' theorem for a continuous event space, the posterior probability is given by the product of the prior probability and the likelihood function (given the evidence *s* and *f*=*n*-*s*), normalized so that the area under the curve equals one, as follows:

$$\begin{aligned}\mathbf{Posterior\ Probability}(x=p|s,n-s) &= \frac{\mathbf{PriorProbability}(x=p;\alpha\mathbf{Prior},\beta\mathbf{Prior})\mathcal{L}(s,f|x=p)}{\int_0^1 \mathbf{PriorProbability}(x=p;\alpha\mathbf{Prior},\beta\mathbf{Prior})\mathcal{L}(s,f|x=p)dx} \\ &= \frac{\binom{n}{s}x^{s+\alpha\mathbf{Prior}-1}(1-x)^{n-s+\beta\mathbf{Prior}-1}/\mathrm{B}(\alpha\mathbf{Prior},\beta\mathbf{Prior})}{\int_0^1 \left(\binom{n}{s}x^{s+\alpha\mathbf{Prior}-1}(1-x)^{n-s+\beta\mathbf{Prior}-1}/\mathrm{B}(\alpha\mathbf{Prior},\beta\mathbf{Prior})\right)dx} \\ &= \frac{x^{s+\alpha\mathbf{Prior}-1}(1-x)^{n-s+\beta\mathbf{Prior}-1}}{\int_0^1 \left(x^{s+\alpha\mathbf{Prior}-1}(1-x)^{n-s+\beta\mathbf{Prior}-1}\right)dx} \\ &= \frac{x^{s+\alpha\mathbf{Prior}-1}(1-x)^{n-s+\beta\mathbf{Prior}-1}}{\mathrm{B}(s+\alpha\mathbf{Prior},n-s+\beta\mathbf{Prior})}.\end{aligned}$$



Posterior Beta densities with samples having success="s", failure="f" of s/(s+f)=1/4, and s+f={3,10,50}, based on 3 different prior probability functions: Haldane (Beta(0,0), Jeffreys (Beta(1/2,1/2)) and Bayes (Beta(1,1)). The image shows that there is little difference between the priors for the posterior with sample size of 50 (with more pronounced peak near *p* = 1/4). Significant differences appear for very small sample sizes (the very skewed distribution for the degenerate case of sample size=3, in this degenerate and unlikely case the Haldane prior results in a reverse "J" shape with mode at *p* = 0 instead of *p*=1/4. If there is sufficient sampling data, the three priors of Bayes (Beta(1,1)), Jeffreys (Beta(1/2,1/2)) and Haldane (Beta(0,0)) should yield similar *posterior* probability densities.

The binomial coefficient

$$\binom{s+f}{s} = \binom{n}{s} = \frac{(s+f)!}{s!f!} = \frac{n!}{s!(n-s)!}$$

appears both in the numerator and the denominator of the posterior probability, and it does not depend on the integration variable x, hence it cancels out, and it is irrelevant to the final result. Similarly the normalizing factor for the prior probability, the beta function B(αPrior,βPrior) cancels out and it is immaterial to the final result. The same posterior probability result can be obtained if one uses an un-normalized prior

$$x^{\alpha\text{Prior}-1}(1-x)^{\beta\text{Prior}-1}$$

because the normalizing factors all cancel out. Several authors (including Jeffreys himself) thus use an un-normalized prior formula since the normalization constant cancels out. The numerator of the posterior probability ends up being just the (un-normalized) product of the prior probability and the likelihood function, and the denominator is its integral from zero to one. The beta function in the denominator, B(s + αPrior, n - s + βPrior), appears as a normalization constant to ensure that the total posterior probability integrates to unity.

The ratio s/n of the number of successes to the total number of trials is a sufficient statistic in the binomial case, which is relevant for the following results.

For the **Bayes'** prior probability (Beta(1,1)), the posterior probability is:

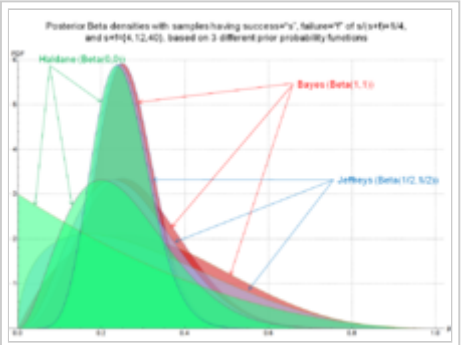
$$\textbf{Posterior Probability}(p=x|s,f) = \frac{x^s(1-x)^{n-s}}{\mathrm{B}(s+1,n-s+1)}, \textbf{ with mean} = \frac{s+1}{n+2}, \textbf{ (and mode= } \frac{s}{n} \textbf{ if } 0<s<n).$$

For the **Jeffreys'** prior probability (Beta(1/2,1/2)), the posterior probability is:

$$\textbf{Posterior Probability}(p=x|s,f) = \frac{x^{s-\frac{1}{2}}(1-x)^{n-s-\frac{1}{2}}}{\mathrm{B}(s+\frac{1}{2},n-s+\frac{1}{2})}, \textbf{ with mean} = \frac{s+\frac{1}{2}}{n+1}, \textbf{ (and mode= } \frac{s-\frac{1}{2}}{n-1} \textbf{ if } \frac{1}{2}<s<n-\frac{1}{2}).$$

and for the **Haldane** prior probability (Beta(0,0)), the posterior probability is:

$$\textbf{Posterior Probability}(p=x|s,f) = \frac{x^{s-1}(1-x)^{n-s-1}}{\mathrm{B}(s,n-s)}, \textbf{ with mean} = \frac{s}{n}, \textbf{ (and mode= } \frac{s-1}{n-2} \textbf{ if } 1<s<n-1).$$



Posterior Beta densities with samples having success="s", failure="f" of s/(s+f)=1/4, and s+f={4,12,40}, based on 3 different prior probability functions: Haldane (Beta(0,0), Jeffreys (Beta(1/2,1/2)) and Bayes (Beta(1,1)). The image shows that there is little difference between the priors for the posterior with sample size of 40 (with more pronounced peak near p=1/4). Significant differences appear for very small sample sizes

From the above expressions it follows that for (s/n)=(1/2) all the above three prior probabilities result in the identical location for the posterior probability mean=mode=1/2. For (s/n)<(1/2), the mean of the posterior probabilities, using the following priors, are such that: mean for Bayes prior > mean for Jeffreys prior > mean for Haldane prior. For (s/n)>(1/2) the order of these inequalities is reversed such that the Haldane prior probability results in the largest posterior mean. The *Haldane* prior probability Beta(0,0) results in a posterior probability density with *mean* (the expected value for the probability of success in the "next" trial) identical to the ratio s/n of the number of successes to the total number of trials. Therefore, the Haldane prior results in a posterior probability with expected value in the next trial equal to the maximum likelihood. The *Bayes* prior probability Beta(1,1) results in a posterior probability density with *mode* identical to the ratio s/n (the maximum likelihood).

In the case that 100% of the trials have been successful (s=n), the *Bayes* prior probability Beta(1,1) results in a posterior expected value equal to the rule of succession (n+1)/(n+2), while the Haldane prior Beta(0,0) results in a posterior expected value of 1 (absolute certainty of success in the next trial). Jeffreys prior probability results in a posterior expected value equal to (n + 1/2)/(n+1), Perks[62] (p. 303) points out: "This provides a new rule of succession and expresses a 'reasonable' position to take up, namely, that after an unbroken run of n successes we assume a probability for the next trial equivalent to the assumption that we are about half-way through an average run, i.e. that we expect a failure once in (2n + 2) trials. The Bayes-Laplace rule implies that we are about at the end of an average run or that we expect a failure once in (n + 2) trials. The comparison clearly favours the new result (what is now called Jeffreys prior) from the point of view of 'reasonableness'."

Conversely, in the case that 100% of the trials have resulted in failure (s=0), the *Bayes* prior probability Beta(1,1) results in a posterior expected value for success in the next trial equal to 1/(n+2), while the Haldane prior Beta(0,0) results in a posterior expected value of success in the next trial of 0 (absolute certainty of failure in the next trial). Jeffreys prior probability results in a posterior expected value for success in the next trial equal to (1/2)/(n+1), which Perks[62] (p. 303) points out: "is a much more reasonably remote result than the Bayes-Laplace result 1/(n + 2)".

Jaynes[50] questions (for the uniform prior Beta(1,1)) the use of these formulas for the cases s=0 or s=n because the integrals do not converge (Beta(1,1) is an improper prior for s=0 or s=n). In practice, the conditions 0<s<n necessary for a mode to exist between both ends for the Bayes prior are usually met, and therefore the Bayes prior (as long as 0<s<n) results in a posterior mode located between both ends of the domain.

As remarked in the section on the rule of succession, K. Pearson showed that after n successes in n trials the posterior probability (based on the Bayes Beta(1,1) distribution as the prior probability) that the next (n + 1) trials will all be successes is exactly 1/2, whatever the value of n. Based on the Haldane Beta(0,0) distribution as the prior probability, this posterior probability is 1 (absolute certainty that after n successes in n trials the next (n + 1) trials will all be successes). Perks[62] (p. 303) shows that, for what is now known as the Jeffreys prior, this probability is ((n + 1/2)/(n+1))((n + 3/2)/(n+2))...(2n - 1/2)/(2n), which for n=1,2,3 gives 3/4, 35/48, 693/960; rapidly approaching a limiting value of $\frac{1}{\sqrt{2}}=0.70710678...$ as n tends to infinity. Perks remarks that what is now known as the Jeffreys prior: "is clearly more 'reasonable' than either the Bayes-Laplace result or the result on the (Haldane) alternative rule rejected by Jeffreys which gives certainty as the probability. It clearly provides a very much better correspondence with the process of induction. Whether it is 'absolutely' reasonable for the purpose, i.e. whether it is yet large enough, without the absurdity of reaching unity, is a matter for others to decide. But it must be realized that the result depends on the assumption of complete indifference and absence of knowledge prior to the sampling experiment."

Following are the variances of the posterior distribution obtained with these three prior probability distributions:

for the **Bayes'** prior probability (Beta(1,1)), the posterior variance is:

$$\text{var} = \frac{(n-s+1)(s+1)}{(3+n)(2+n)^2}, \text{ which for } s = \frac{n}{2} \text{ results in var} = \frac{1}{12+4n}$$

for the **Jeffreys'** prior probability (Beta(1/2,1/2)), the posterior variance is:

$$\text{var} = \frac{(n-s+\frac{1}{2})(s+\frac{1}{2})}{(2+n)(1+n)^2}, \text{ which for } s = \frac{n}{2} \text{ results in } \text{var} = \frac{1}{8+4n}$$

and for the **Haldane** prior probability (Beta(0,0)), the posterior variance is:

$$\text{var} = \frac{(n-s)s}{(1+n)n^2}, \text{ which for } s = \frac{n}{2} \text{ results in var} = \frac{1}{4+4n}$$

So, as remarked by Silvey,[48] for large *n*, the variance is small and hence the posterior distribution is highly concentrated, whereas the assumed prior distribution was very diffuse. This is in accord with what one would hope for, as vague prior knowledge is transformed (through Bayes theorem) into a more precise posterior knowledge by an informative experiment. For small *n* the Haldane Beta(0,0) prior results in the largest posterior variance while the Bayes Beta(1,1) prior results in the more concentrated posterior. Jeffreys prior Beta(1/2,1/2) results in a posterior variance in between the other two. As *n* increases, the variance rapidly decreases so that the posterior variance for all three priors converges to approximately the same value (approaching zero variance as $n \to \infty$). Recalling the previous result that the *Haldane* prior probability Beta(0,0) results in a posterior probability density with *mean* (the expected value for the probability of success in the "next" trial) identical to the ratio s/n of the number of successes to the total number of trials, it follows from the above expression that also the *Haldane* prior Beta(0,0) results in a posterior with *variance* identical to the variance expressed in terms of the max. likelihood estimate s/n and sample size (in section titled "Variance"):

$$\text{var} = \frac{\mu(1-\mu)}{1+\nu} = \frac{(n-s)s}{(1+n)n^2}$$

with the mean $\mu = s/n$ and the sample size $\nu = n$.

In Bayesian inference, using a prior distribution Beta(αPrior,βPrior) prior to a binomial distribution is equivalent to adding (αPrior - 1) pseudo-observations of "success" and (βPrior - 1) pseudo-observations of "failure" to the actual number of successes and failures observed, then estimating the parameter *p* of the binomial distribution by the proportion of successes over both real- and pseudo-observations. A uniform prior Beta(1,1) does not add (or subtract) any pseudo-observations since for Beta(1,1) it follows that (αPrior - 1)=0 and (βPrior - 1)=0. The Haldane prior Beta(0,0) subtracts one pseudo observation from each and Jeffreys prior Beta(1/2,1/2) subtracts 1/2 pseudo-observation of success and an equal number of failure. This subtraction has the effect of smoothing out the posterior

distribution. If the proportion of successes is not 50% ($s/n \neq 1/2$) values of αPrior and βPrior less than 1 (and therefore negative (αPrior - 1) and (βPrior - 1)) favor sparsity, i.e. distributions where the parameter p is closer to either 0 or 1. In effect, values of $\alpha Prior$ and $\beta Prior$ between 0 and 1, when operating together, function as a concentration parameter.

The accompanying plots show the posterior probability density functions for sample sizes n={3,10,50}, successes s={n/2,n/4} and Beta(αPrior,βPrior)= {Beta(0,0),Beta(1/2,1/2),Beta(1,1)}. Also shown are the cases for n={4,12,40}, success s={n/4} and Beta(αPrior,βPrior)={Beta(0,0),Beta(1/2,1/2),Beta(1,1)}. The first plot shows the symmetric cases, for successes s={n/2}, with mean=mode=1/2 and the second plot shows the skewed cases s={n/4}. The images show that there is little difference between the priors for the posterior with sample size of 50 (characterized by a more pronounced peak near p=1/2). Significant differences appear for very small sample sizes (in particular for the flatter distribution for the degenerate case of sample size=3). Therefore, the skewed cases, with successes s={n/4}, show a larger effect from the choice of prior, at small sample size, than the symmetric cases. For symmetric distributions, the Bayes prior Beta(1,1) results in the most "peaky" and highest posterior distributions and the Haldane prior Beta(0,0) results in the flattest and lowest peak distribution. The Jeffreys prior Beta(1/2,1/2) lies in between them. For nearly symmetric, not too skewed distributions the effect of the priors is similar. For very small sample size (in this case for a sample size of 3) and skewed distribution (in this example for s={n/4}) the Haldane prior can result in a reverse-J-shaped distribution with a singularity at the left end. However, this happens only in degenerate cases (in this example n=3 and hence s=3/4<1, a degenerate value because s should be greater than unity in order for the posterior of the Haldane prior to have a mode located between the ends, and because s=3/4 is not an integer number, hence it violates the initial assumption of a binomial distribution for the likelihood) and it is not an issue in generic cases of reasonable sample size (such that the condition 1<s<n - 1, necessary for a mode to exist between both ends, is fulfilled).

In Chapter 12 (p. 385) of his book, Jaynes[50] asserts that the *Haldane prior* Beta(0,0) describes a *prior state of knowledge of complete ignorance*, where we are not even sure whether it is physically possible for an experiment to yield either a success or a failure, while the *Bayes (uniform) prior Beta(1,1) applies if* one knows that *both binary outcomes are possible*. Jaynes states: "*interpret the Bayes-Laplace (Beta(1,1)) prior as describing not a state of complete ignorance*, but the state of knowledge in which we have observed one success and one failure...once we have seen at least one success and one failure, then we know that the experiment is a true binary one, in the sense of physical possibility." Jaynes [50] does not specifically discuss Jeffreys prior Beta(1/2,1/2) (Jaynes discussion of "Jeffreys prior" on pp. 181, 423 and on chapter 12 of Jaynes book[50] refers instead to the improper, un-normalized, prior "1/p" introduced by Jeffreys in the 1939 edition of his book,[60] seven years before he introduced what is now known as Jeffreys' invariant prior: the square root of the determinant of Fisher's information matrix. *"1/p" is Jeffreys' (1946) invariant prior for the exponential distribution, not for the Bernoulli or binomial distributions*). However, it follows from the above discussion that Jeffreys Beta(1/2,1/2) prior represents a state of knowledge in between the Haldane Beta(0,0) and Bayes Beta (1,1) prior.

Similarly, Karl Pearson in his 1892 book The Grammar of Science[69][70] (p. 144 of 1900 edition) maintained that the Bayes (Beta(1,1) uniform prior was not a complete ignorance prior, and that it should be used when prior information justified to "distribute our ignorance equally"". K. Pearson wrote: "Yet the only supposition that we appear to have made is this: that, knowing nothing of nature, routine and anomy (from the Greek ανομία, namely: a- "without", and nomos "law") are to be considered as equally likely to occur. Now we were not really justified in making even this assumption, for it involves a knowledge that we do not possess regarding nature. We use our *experience* of the constitution and action of coins in general to assert that heads and tails are equally probable, but we have no right to assert before experience that, as we know nothing of nature, routine and breach are equally probable. In our ignorance we ought to consider before experience that nature may consist of all routines, all anomies (normlessness), or a mixture of the two in any proportion whatever, and that all such are equally probable. Which of these constitutions after experience is the most probable must clearly depend on what that experience has been like."

If there is sufficient sampling data, *and the posterior probability mode is not located at one of the extremes of the domain* (x=0 or x=1), the three priors of Bayes (Beta(1,1)), Jeffreys (Beta(1/2,1/2)) and Haldane (Beta(0,0)) should yield similar *posterior* probability densities. Otherwise, as Gelman et al.[71] (p. 65) point out, "if so few data are available that the choice of noninformative prior distribution makes a difference, one should put relevant information into the prior distribution", or as Berger[16] (p. 125) points out "when different reasonable priors yield substantially different answers, can it be right to state that there *is* a single answer? Would it not be better to admit that there is scientific uncertainty, with the conclusion depending on prior beliefs?."

Subjective logic

In standard logic, propositions are considered to be either true or false. In contradistinction, subjective logic assumes that humans cannot determine with absolute certainty whether a proposition about the real world is absolutely true or false. In subjective logic the posteriori probability estimates of binary events can be represented by beta distributions.[72]

Wavelet analysis

A wavelet is a wave-like oscillation with an amplitude that starts out at zero, increases, and then decreases back to zero. It can typically be visualized as a "brief oscillation" that promptly decays. Wavelets can be used to extract information from many different kinds of data, including – but certainly not limited to – audio signals and images. Thus, wavelets are purposefully crafted to have specific properties that make them useful for signal processing. Wavelets are localized in both time and frequency whereas the standard Fourier transform is only localized in frequency. Therefore, standard Fourier Transforms are only applicable to stationary processes, while wavelets are applicable to non-stationary processes. Continuous wavelets can be constructed based on the beta distribution. Beta wavelets[73] can be viewed as a soft variety of Haar wavelets whose shape is fine-tuned by two shape parameters α and β.

Project management: task cost and schedule modeling

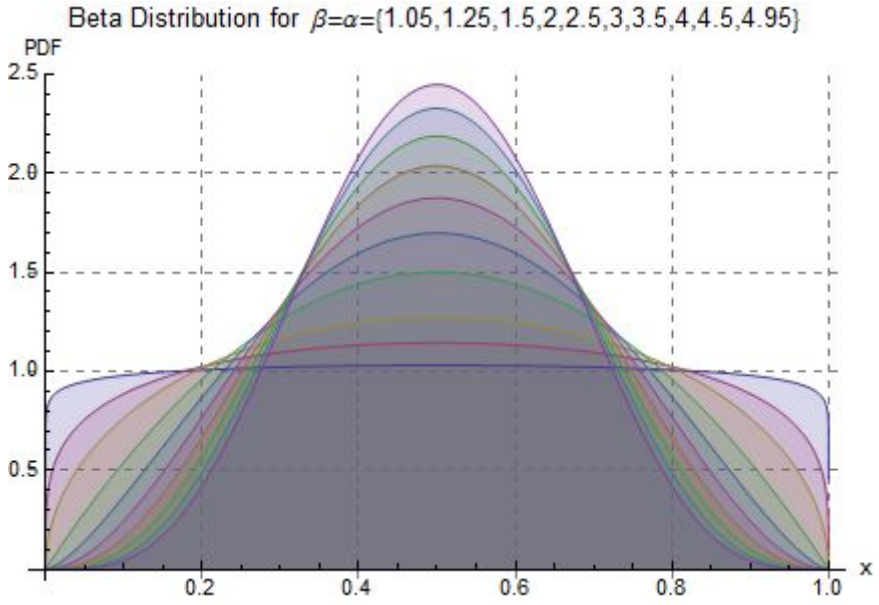
The beta distribution can be used to model events which are constrained to take place within an interval defined by a minimum and maximum value. For this reason, the beta distribution — along with the triangular distribution — is used extensively in PERT, critical path method (CPM), Joint Cost Schedule Modeling (JCSM) and other project management/control systems to describe the time to completion and the cost of a task. In project management, shorthand computations are widely used to estimate the mean and standard deviation of the beta distribution:[2]

$$\mu(X) = \frac{a + 4b + c}{6}$$
$$\sigma(X) = \frac{c - a}{6}$$

where a is the minimum, c is the maximum, and b is the most likely value (the mode for $\alpha > 1$ and $\beta > 1$).

The above estimate for the mean $\mu(X) = \frac{a + 4b + c}{6}$ is known as the PERT three-point estimation and it is exact for either of the following values of β (for arbitrary α within these ranges):

β = α > 1 (symmetric case) with standard deviation $\sigma(X) = \frac{(c - a)}{2\sqrt{1 + 2\alpha}}$, skewness = 0, and excess kurtosis = $\frac{-6}{3 + 2\alpha}$

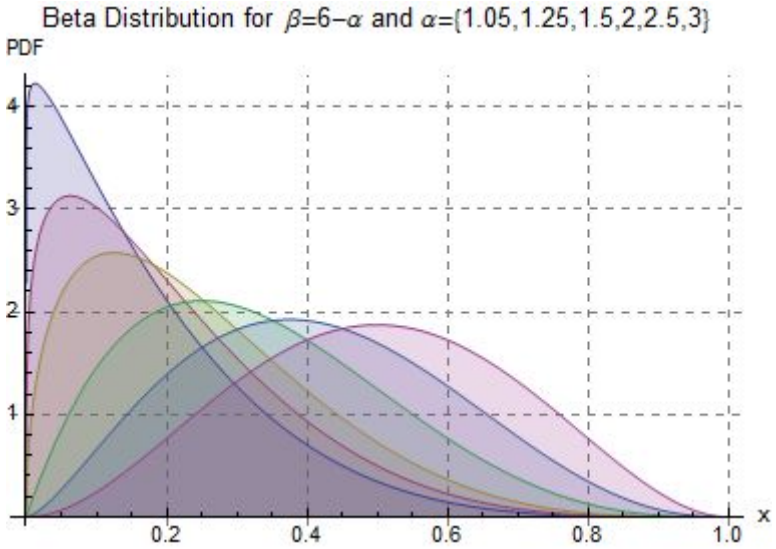


or

β = 6−α for 5 > α > 1 (skewed case) with standard deviation

$$\sigma(X) = \frac{(c - a)\sqrt{\alpha(6 - \alpha)}}{6\sqrt{7}},$$

skewness = $\frac{(3 - \alpha)\sqrt{7}}{2\sqrt{\alpha(6 - \alpha)}}$, and excess kurtosis = $\frac{21}{\alpha(6 - \alpha)} - 3$

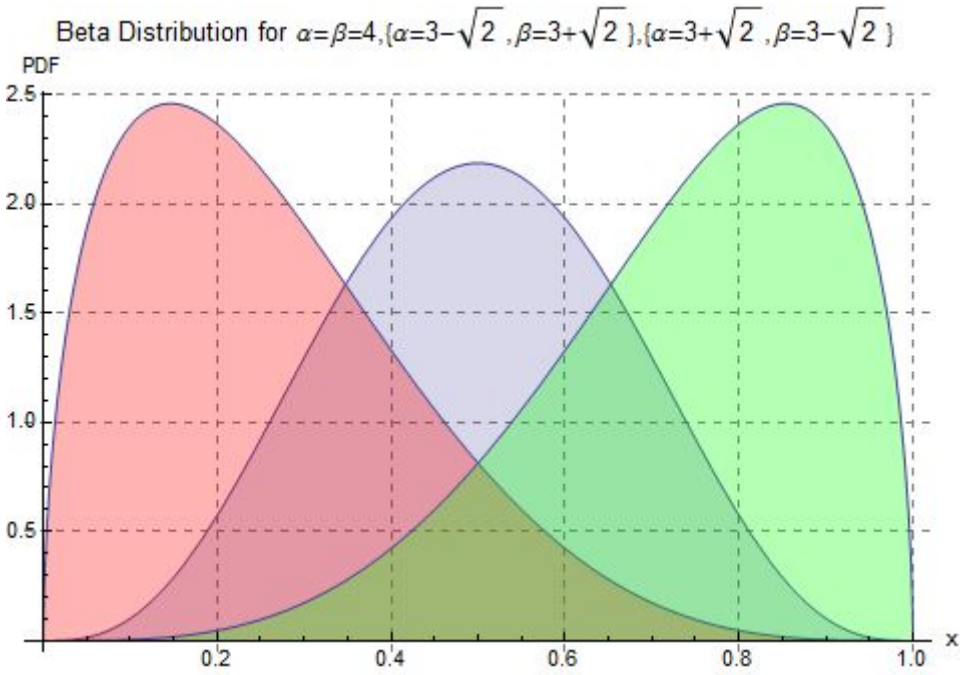


The above estimate for the standard deviation σ(*X*) = (*c*−*a*)/6 is exact for either of the following values of α and β:

α = β = 4 (symmetric) with skewness = 0, and excess kurtosis = −6/11.

β = 6−α and $\boldsymbol{\alpha = 3 - \sqrt{2}}$ (right-tailed, positive skew) with skewness = $\frac{\mathbf{1}}{\sqrt{\mathbf{2}}}$, and excess kurtosis = 0

β = 6−α and $\boldsymbol{\alpha = 3 + \sqrt{2}}$ (left-tailed, negative skew) with skewness = $\frac{\mathbf{-1}}{\sqrt{\mathbf{2}}}$, and excess kurtosis = 0



Otherwise, these can be poor approximations for beta distributions with other values of α and β, exhibiting average errors of 40% in the mean and 549% in the variance.[74][75][76]

Alternative parametrizations

Two parameters

Mean and sample size

The beta distribution may also be reparameterized in terms of its mean μ ($0 < \mu < 1$) and the addition of both shape parameters $\nu = \alpha + \beta > 0$([15] p. 83). Denoting by αPosterior and βPosterior the shape parameters of the posterior beta distribution resulting from applying Bayes theorem to a binomial likelihood function and a prior probability, the interpretation of the addition of both shape parameters to be sample size = ν = αPosterior + βPosterior is only correct for the Haldane prior probability Beta(0,0). Specifically, for the Bayes (uniform) prior Beta(1,1) the correct interpretation would be sample size = αPosterior + βPosterior − 2, or ν = (sample size) + 2. Of course, for sample size much larger than 2, the difference between these two priors becomes negligible. (See section Bayesian inference for further details.) In the rest of this article $\nu = \alpha + \beta$ will be referred to as "sample size", but one should remember that it is, strictly speaking, the "sample size" of a binomial likelihood function only when using a Haldane Beta(0,0) prior in Bayes theorem.

This parametrization may be useful in Bayesian parameter estimation. For example, one may administer a test to a number of individuals. If it is assumed that each person's score ($0 \le \theta \le 1$) is drawn from a population-level Beta distribution, then an important statistic is the mean of this population-level distribution. The mean and sample size parameters are related to the shape parameters α and β via[15]

$$\alpha = \mu\nu,\ \beta = (1-\mu)\nu$$

Under this parametrization, one may place an uninformative prior probability over the mean, and a vague prior probability (such as an exponential or gamma distribution) over the positive reals for the sample size, if they are independent, and prior data and/or beliefs justify it.

Mode and concentration

The mode and "concentration" $\kappa = \alpha + \beta$ can also be used to calculate the parameters for a beta distribution.[77]

$$\alpha = \omega(\kappa - 2) + 1$$
$$\beta = (1 - \omega)(\kappa - 2) + 1$$

Mean (allele frequency) and (Wright's) genetic distance between two populations

The Balding–Nichols model[1] is a two-parameter parametrization of the beta distribution used in population genetics. It is a statistical description of the allele frequencies in the components of a sub-divided population:

$$\alpha = \mu\nu,$$
$$\beta = (1 - \mu)\nu,$$

where $\nu = \alpha + \beta = \frac{1 - F}{F}$ and $0 < F < 1$; here F is (Wright's) genetic distance between two populations.

See the articles Balding–Nichols model, F-statistics, fixation index and coefficient of relationship, for further information.

Mean and variance

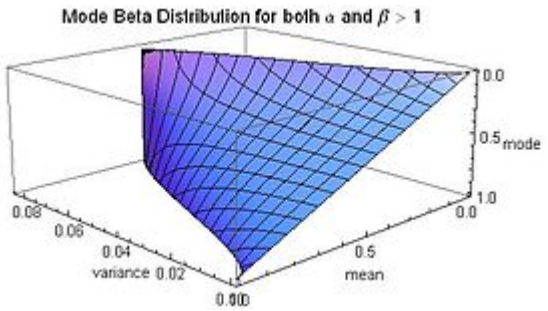
Solving the system of (coupled) equations given in the above sections as the equations for the mean and the variance of the beta distribution in terms of the original parameters α and β, one can express the α and β parameters in terms of the mean (μ) and the variance (var):

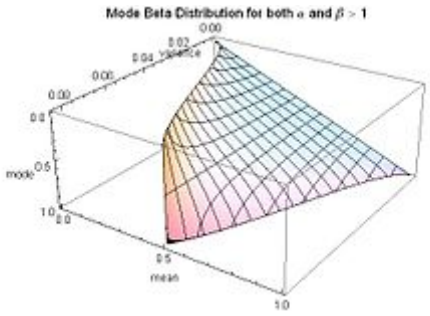
$$\nu = \alpha + \beta = \frac{\mu(1-\mu)}{\mathrm{var}} - 1, \text{ where } \nu = (\alpha+\beta) > 0, \text{ therefore: } \mathrm{var} < \mu(1-\mu)$$

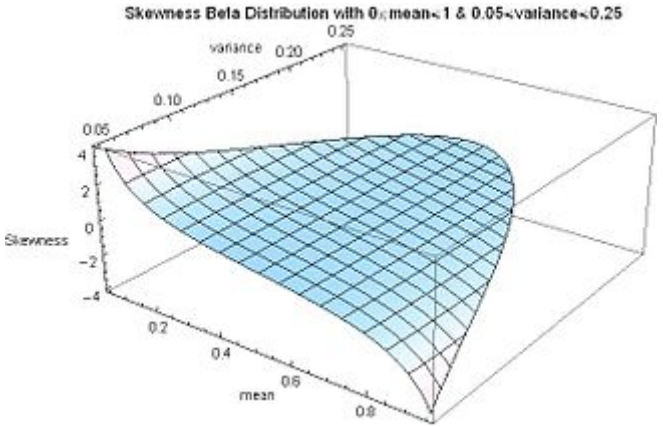
$$\alpha = \mu\nu = \mu\left(\frac{\mu(1-\mu)}{\mathrm{var}} - 1\right), \text{ if } \mathrm{var} < \mu(1-\mu)$$

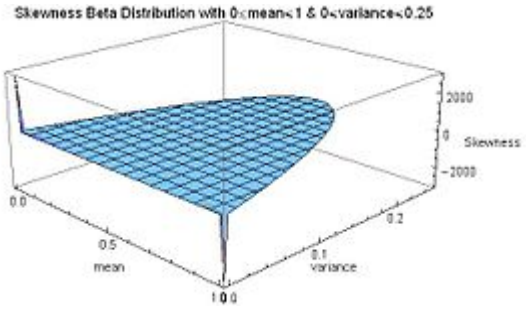
$$\beta = (1-\mu)\nu = (1-\mu)\left(\frac{\mu(1-\mu)}{\mathrm{var}} - 1\right), \text{ if } \mathrm{var} < \mu(1-\mu).$$

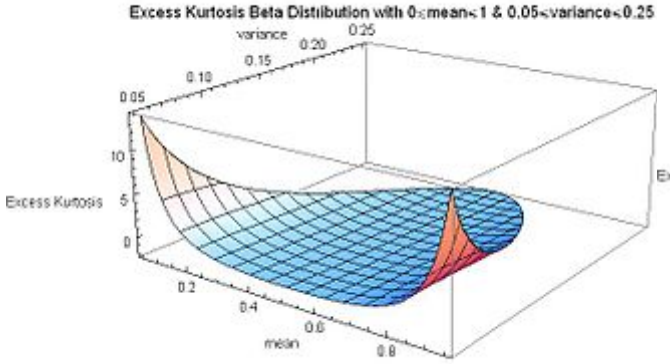
This parametrization of the beta distribution may lead to a more intuitive understanding than the one based on the original parameters α and β. For example, by expressing the mode, skewness, excess kurtosis and differential entropy in terms of the mean and the variance:

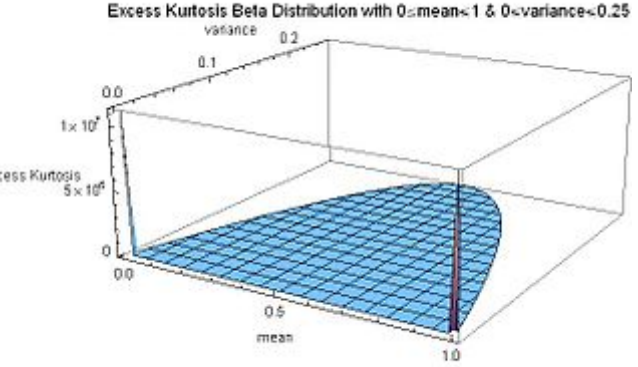


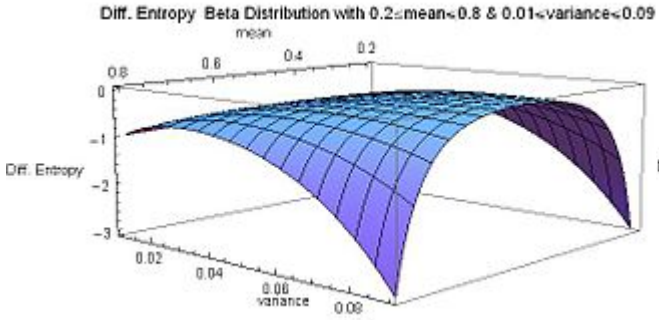


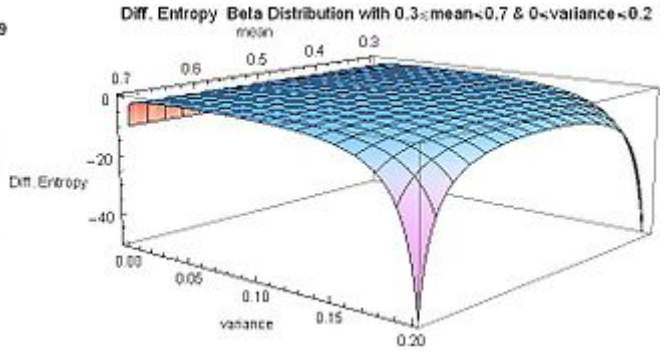












Four parameters

A beta distribution with the two shape parameters α and β is supported on the range [0,1] or (0,1). It is possible to alter the location and scale of the distribution by introducing two further parameters representing the minimum, *a*, and maximum *c* (*c* > *a*), values of the distribution,[7] by a linear transformation substituting the non-dimensional variable *x* in terms of the new variable *y* (with support [*a*,*c*] or (*a*,*c*)) and the parameters *a* and *c*:

$$y = x(c-a) + a,\ \text{therefore}\ x = \frac{y-a}{c-a}.$$

The probability density function of the four parameter beta distribution is equal to the two parameter distribution, scaled by the range (*c*-*a*), (so that the total area under the density curve equals a probability of one), and with the "y" variable shifted and scaled as follows:

$$f(y;\alpha,\beta,a,c) = \frac{f(x;\alpha,\beta)}{c-a} = \frac{\left(\frac{y-a}{c-a}\right)^{\alpha-1}\left(\frac{c-y}{c-a}\right)^{\beta-1}}{(c-a)B(\alpha,\beta)} = \frac{(y-a)^{\alpha-1}(c-y)^{\beta-1}}{(c-a)^{\alpha+\beta-1}B(\alpha,\beta)}.$$

That a random variable *Y* is Beta-distributed with four parameters α, β, *a*, and *c* will be denoted by:

$$Y \sim Beta(\alpha,\beta,a,c).$$

The measures of central location are scaled (by (*c*-*a*)) and shifted (by *a*), as follows:

$$\begin{aligned}
\operatorname{mean}(Y) &= \operatorname{mean}(X)(c-a) + a = \left(\frac{\alpha}{\alpha+\beta}\right)(c-a) + a = \frac{\alpha c + \beta a}{\alpha+\beta} \\
\operatorname{mode}(Y) &= \operatorname{mode}(X)(c-a) + a = \left(\frac{\alpha-1}{\alpha+\beta-2}\right)(c-a) + a = \frac{(\alpha-1)c + (\beta-1)a}{\alpha+\beta-2}, \qquad \text{if } \alpha,\beta > 1 \\
\operatorname{median}(Y) &= \operatorname{median}(X)(c-a) + a = \left(I^{[-1]}_{\frac{1}{2}}(\alpha,\beta)\right)(c-a) + a \\
G_Y &= G_X(c-a) + a = \left(e^{\psi(\alpha)-\psi(\alpha+\beta)}\right)(c-a) + a \\
H_Y &= H_X(c-a) + a = \left(\frac{\alpha-1}{\alpha+\beta-1}\right)(c-a) + a, \qquad \text{if } \alpha,\beta > 0
\end{aligned}$$

The statistical dispersion measures are scaled (they do not need to be shifted because they are already centered on the mean) by the range (c-a), linearly for the mean deviation and nonlinearly for the variance:

$$\begin{aligned}
&(\text{mean deviation around mean})(Y) = \\
&((\text{mean deviation around mean})(X))(c-a) = \frac{2\alpha^\alpha\beta^\beta}{\mathrm{B}(\alpha,\beta)(\alpha+\beta)^{\alpha+\beta+1}}(c-a)
\end{aligned}$$

$$\operatorname{var}(Y) = \operatorname{var}(X)(c-a)^2 = \frac{\alpha\beta(c-a)^2}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Since the skewness and excess kurtosis are non-dimensional quantities (as moments centered on the mean and normalized by the standard deviation), they are independent of the parameters *a* and *c*, and therefore equal to the expressions given above in terms of *X* (with support [0,1] or (0,1)):

$$\operatorname{skewness}(Y) = \operatorname{skewness}(X) = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}.$$

$$\operatorname{kurtosis\ excess}(Y) = \operatorname{kurtosis\ excess}(X) = \frac{6[(\alpha-\beta)^2(\alpha+\beta+1) - \alpha\beta(\alpha+\beta+2)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}$$

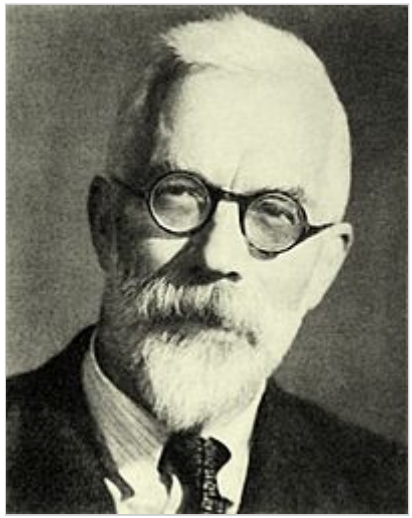
History

The first systematic modern discussion of the beta distribution is probably due to Karl Pearson FRS[78] (27 March 1857 – 27 April 1936[79]), an influential English mathematician who has been credited with establishing the discipline of mathematical statistics.[80] In Pearson's papers[27][38] the beta distribution is couched as a solution of a differential equation: Pearson's Type I distribution which it is essentially identical to except for arbitrary shifting and re-scaling (the beta and Pearson Type I distributions can always be equalized by proper choice of parameters). In fact, in several English books and journal articles in the few decades prior to World War II, it was common to refer to the beta distribution as Pearson's Type I distribution. William P. Elderton (1877–1962) in his 1906 monograph "Frequency curves and correlation"[40] further analyzes the beta distribution as Pearson's Type I distribution, including a full discussion of the method of moments for the four parameter case, and diagrams of (what Elderton describes as) U-shaped, J-shaped, twisted J-shaped, "cocked-hat" shapes, horizontal and angled straight-line cases. Elderton wrote "I am chiefly indebted to Professor Pearson, but the indebtedness is of a kind for which it is impossible to offer formal thanks." Elderton in his 1906 monograph [40] provides an impressive amount of information on the beta distribution, including equations for the origin of the distribution chosen to be the mode, as well as for other Pearson distributions: types I through VII. Elderton also included a number of appendixes, including one appendix ("II") on the beta and gamma functions. In later editions, Elderton added equations for the origin of the distribution chosen to be the mean, and analysis of Pearson distributions VIII through XII.

Karl Pearson analyzed the beta distribution as the solution Type I of Pearson distributions

As remarked by Bowman and Shenton [42] "Fisher and Pearson had a difference of opinion in the approach to (parameter) estimation, in particular relating to (Pearson's method of) moments and (Fisher's method of) maximum likelihood in the case of the Beta distribution." Also according to Bowman and Shenton, "the case of a Type I (beta distribution) model being the center of the controversy was pure serendipity. A more difficult model of 4 parameters would have been hard to find." Ronald Fisher (17 February 1890 – 29 July 1962) was one of the giants of

statistics in the first half of the 20th century, and his long running public conflict with Karl Pearson can be followed in a number of articles in prestigious journals. For example, concerning the estimation of the four parameters for the beta distribution, and Fisher's criticism of Pearson's method of moments as being arbitrary, see Pearson's article "Method of moments and method of maximum likelihood" [43] (published three years after his retirement from University College, London, where his position had been divided between Fisher and Pearson's son Egon) in which Pearson writes "I read (Koshai's paper in the Journal of the Royal Statistical Society, 1933) which as far as I am aware is the only case at present published of the application of Professor Fisher's method. To my astonishment that method depends on first working out the constants of the frequency curve by the (Pearson) Method of Moments and then superposing on it, by what Fisher terms "the Method of Maximum Likelihood" a further approximation to obtain, what he holds, he will thus get, "more efficient values" of the curve constants."



Biologist and statistician Ronald Fisher

David and Edwards's treatise on the history of statistics[81] cites the first modern treatment of the beta distribution, in 1911,[82] using the beta designation that has become standard, due to Corrado Gini,(May 23, 1884 – March 13, 1965), an Italian statistician, demographer, and sociologist, who developed the Gini coefficient. N.L.Johnson and S.Kotz, in their comprehensive and very informative monograph[83] on leading historical personalities in statistical sciences credit Corrado Gini[84] as "an early Bayesian...who dealt with the problem of eliciting the parameters of an initial Beta distribution, by singling out techniques which anticipated the advent of the so called empirical Bayes approach." Bayes, in a posthumous paper [63] published in 1763 by Richard Price, obtained a beta distribution as the density of the probability of success in Bernoulli trials (see the section titled "Applications, Bayesian inference" in this article), but the paper does not analyze any of the moments of the beta distribution or discuss any of its properties.

References

1. Balding, David J.; Nichols, Richard A. (1995). "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity". *Genetica*. Springer. **96** (1–2): 3–12. doi:10.1007/BF01441146. PMID 7607457.
2. Malcolm, D. G.; Roseboom, J. H.; Clark, C. E.; Fazar, W. (September–October 1958). "Application of a Technique for Research and Development Program Evaluation". *Operations Research*. **7** (5): 646–669. doi:10.1287/opre.7.5.646. ISSN 0030-364X.
3. Sulaiman, M.Yusof; Oo, W. M. Hlaing; Wahab, Mahdi Abd; Zakaria, Azmi (December 1999). "Application of beta distribution model to Malaysian sunshine data". *Renewable Energy*. **18** (4): 573–579. doi:10.1016/S0960-1481(99)00002-6.
4. Haskett, Jonathan D.; Pachepsky, Yakov A.; Acock, Basil (1995). "Use of the beta distribution for parameterizing variability of soil properties at the regional level for crop yield estimation". *Agricultural Systems*. **48** (1): 73–86. doi:10.1016/0308-521X(95)93646-U.
5. Gullco, Robert S.; Anderson, Malcolm (December 2009). "Use of the Beta Distribution To Determine Well-Log Shale Parameters". *SPE Reservoir Evaluation & Engineering*. **12** (6): 929–942. doi:10.2118/106746-PA.
6. Wiley, James A.; Herschkorn, Stephen J.; Padian, Nancy S. (January 1989). "Heterogeneity in the probability of HIV transmission per sexual contact: The case of male-to-female transmission in penile—vaginal intercourse". *Statistics in Medicine*. **8** (1): 93–102. doi:10.1002/sim.4780080110.
7. Johnson, Norman L.; Kotz, Samuel; Balakrishnan, N. (1995). "Chapter 21:Beta Distributions". *Continuous Univariate Distributions Vol. 2* (2nd ed.). Wiley. ISBN 978-0-471-58494-0.
8. Keeping, E. S. (2010). *Introduction to Statistical Inference*. Dover Publications. ISBN 978-0486685021.

9. Wadsworth, George P. and Joseph Bryan (1960). *Introduction to Probability and Random Variables*. McGraw-Hill.
10. Hahn, Gerald J.; Shapiro, S. (1994). *Statistical Models in Engineering (Wiley Classics Library)*. Wiley-Interscience. ISBN 978-0471040651.
11. Feller, William (1971). *An Introduction to Probability Theory and Its Applications, Vol. 2*. Wiley. ISBN 978-0471257097.
12. Gupta (Editor), Arjun K. (2004). *Handbook of Beta Distribution and Its Applications*. CRC Press. ISBN 978-0824753962.
13. Panik, Michael J (2005). *Advanced Statistics from an Elementary Point of View*. Academic Press. ISBN 978-0120884940.
14. Rose, Colin; Smith, Murray D. (2002). *Mathematical Statistics with MATHEMATICA*. Springer. ISBN 978-0387952345.
15. Kruschke, John K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. p. 83: Academic Press / Elsevier. ISBN 978-0123814852.
16. Berger, James O. (2010). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer. ISBN 978-1441930743.
17. Kerman J (2011) "A closed-form approximation for the median of the beta distribution". arXiv:1111.0433v1
18. Mosteller, Frederick and John Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley Pub. Co.,. ISBN 978-0201048544.
19. Feller, William (1968). *An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd Edition*. ISBN 978-0471257080.
20. Philip J. Fleming and John J. Wallace. *How not to lie with statistics: the correct way to summarize benchmark results*. Communications of the ACM, 29(3):218–221, March 1986.
21. "NIST/SEMATECH e-Handbook of Statistical Methods 1.3.6.6.17. Beta Distribution". *National Institute of Standards and Technology Information Technology Laboratory*. April 2012. Retrieved May 31, 2016.
22. Oguamanam, D.C.D.; Martin, H. R.; Huissoon, J. P. (1995). "On the application of the beta distribution to gear damage analysis". *Applied Acoustics*. **45** (3): 247–261. doi:10.1016/0003-682X(95)00001-P.
23. Zhiqiang Liang; Jianming Wei; Junyu Zhao; Haitao Liu; Baoqing Li; Jie Shen; Chunlei Zheng (27 August 2008). "The Statistical Meaning of Kurtosis and Its New Application to Identification of Persons Based on Seismic Signals". *Sensors*. **8**: 5106–5119. doi:10.3390/s8085106.
24. Kenney, J. F., and E. S. Keeping (1951). *Mathematics of Statistics Part Two, 2nd edition*. D. Van Nostrand Company Inc.
25. Abramowitz, Milton and Irene A. Stegun (1965). *Handbook Of Mathematical Functions With Formulas, Graphs, And Mathematical Tables*. Dover. ISBN 978-0-486-61272-0.
26. Weisstein., Eric W. "Kurtosis". MathWorld--A Wolfram Web Resource. Retrieved 13 August 2012.
27. Pearson, Karl (1916). "Mathematical contributions to the theory of evolution, XIX: Second supplement to a memoir on skew variation". *Philosophical Transactions of the Royal Society A*. **216** (538–548): 429–457. Bibcode:1916RSPTA.216..429P. doi:10.1098/rsta.1916.0009. JSTOR 91092.
28. Gradshteyn, Izrail Solomonovich; Ryzhik, Iosif Moiseevich; Geronimus, Yuri Veniaminovich; Tseytlin, Michail Yulyevich; Jeffrey, Alan (2015) [October 2014]. Zwillinger, Daniel; Moll, Victor Hugo, eds. *Table of Integrals, Series, and Products*. Translated by Scripta Technica, Inc. (8 ed.). Academic Press, Inc. ISBN 0-12-384933-0. LCCN 2014010276. ISBN 978-0-12-384933-5.
29. Billingsley, Patrick (1995). "30". *Probability and measure* (3rd ed.). Wiley-Interscience. ISBN 0-471-00710-2.
30. MacKay, David (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press; First Edition. ISBN 978-0521642989.
31. Johnson, N.L. (1949). "Systems of frequency curves generated by methods of translation". *Biometrika*. **36**: 149–176. doi:10.1093/biomet/36.1-2.149.
32. A. C. G. Verdugo Lazo and P. N. Rathie. "On the entropy of continuous probability distributions," *IEEE Trans. Inf. Theory*, IT-24:120–122, 1978.
33. Shannon, Claude E., "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27 (4):623–656,1948.PDF (http://www.alcatel-lucent.com/bstj/vol27-1948/articles/bstj27-4-623.pdf)
34. Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience; 2 edition. ISBN 978-0471241959.
35. Plunkett, Kim, and Jeffrey Elman (1997). *Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations (Neural Network Modeling and Connectionism)*. p. 166: A Bradford Book. ISBN 978-0262661058.
36. Nallapati, Ramesh (2006). *The smoothed dirichlet distribution: understanding cross-entropy ranking in information retrieval*. Ph.D. thesis: Computer Science Dept., University of Massachusetts Amherst.
37. Pearson, Egon S. (July 1969). "Some historical reflections traced through the development of the use of frequency curves". *THEMIS Statistical Analysis Research Program, Technical Report 38*. Office of Naval Research, Contract N000014-68-A-0515 (Project NR 042-260).

38. Pearson, Karl (1895). "Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material". *Philosophical Transactions of the Royal Society*. **186**: 343–414. Bibcode:1895RSPTA.186..343P. doi:10.1098/rsta.1895.0010. JSTOR 90649.
39. Engineering Statistics Handbook (http://www.itl.nist.gov/div898/handbook/eda/section3/eda366h.htm)
40. Elderton, William Palin (1906). *Frequency-Curves and Correlation* (PDF). Charles and Edwin Layton (London).
41. Elderton, William Palin and Norman Lloyd Johnson (2009). *Systems of Frequency Curves*. Cambridge University Press. ISBN 978-0521093361.
42. Bowman, K. O.; Shenton, L. R. (2007). "The beta distribution, moment method, Karl Pearson and R.A. Fisher" (PDF). *Far East J. Theo. Stat*. **23** (2): 133–164.
43. Pearson, Karl (June 1936). "Method of moments and method of maximum likelihood". *Biometrika*. **28** (1/2): 34. doi:10.2307/2334123.
44. Joanes, D. N.; C. A. Gill (1998). "Comparing measures of sample skewness and kurtosis". *The Statistician*. **47** (Part 1): 183–189. doi:10.1111/1467-9884.00122.
45. Beckman, R. J.; G. L. Tietjen (1978). "Maximum likelihood estimation for the beta distribution". *Journal of Statistical Computation and Simulation*. **7** (3-4): 253–258. doi:10.1080/00949657808810232.
46. Gnanadesikan, R.,Pinkham and Hughes (1967). "Maximum likelihood estimation of the parameters of the beta distribution from smallest order statistics". *Technometrics*. **9**: 607–620. doi:10.2307/1266199.
47. Fackler, Paul. "Inverse Digamma Function (Matlab)". Harvard University School of Engineering and Applied Sciences. Retrieved 2012-08-18.
48. Silvey, S.D. (1975). *Statistical Inference*. page 40: Chapman and Hal. ISBN 978-0412138201.
49. Edwards, A. W. F. (1992). *Likelihood*. The Johns Hopkins University Press. ISBN 978-0801844430.
50. Jaynes, E.T. (2003). *Probability theory, the logic of science*. Cambridge University Press. ISBN 978-0521592710.
51. Costa, Max, and Cover, Thomas (September 1983). *On the similarity of the entropy power inequality and the Brunn Minkowski inequality* (PDF). Tech.Report 48, Dept. Statistics, Stanford University.
52. Aryal, Gokarna; Saralees Nadarajah (2004). "Information matrix for beta distributions" (PDF). *Serdica Mathematical Journal (Bulgarian Academy of Science)*. **30**: 513–526.
53. van der Waerden, B. L., "Mathematical Statistics", Springer, ISBN 978-3-540-04507-6.
54. David, H. A., Nagaraja, H. N. (2003) *Order Statistics* (3rd Edition). Wiley, New Jersey pp 458. ISBN 0-471-38926-9
55. Herrerías-Velasco, José Manuel and Herrerías-Pleguezuelo, Rafael and René van Dorp, Johan. (2011). Revisiting the PERT mean and Variance. European Journal of Operational Research (210), p. 448–451.
56. Laplace, Pierre Simon, marquis de (1902). *A philosophical essay on probabilities*. New York : J. Wiley ; London : Chapman & Hall. ISBN 1-60206-328-1.
57. Cox, Richard T. (1961). *Algebra of Probable Inference*. The Johns Hopkins University Press. ISBN 978-0801869822.
58. Keynes, John Maynard (2010) [1921]. *A Treatise on Probability: The Connection Between Philosophy and the History of Science*. Wildside Press. ISBN 978-1434406965.
59. Pearson, Karl (1907). "On the Influence of Past Experience on Future Expectation". *Philosophical Magazine*. **6** (13): 365–378.
60. Jeffreys, Harold (1998). *Theory of Probability*. Oxford University Press, 3rd edition. ISBN 978-0198503682.
61. Broad, C. D. (October 1918). "On the relation between induction and probability". *MIND, a quarterly review of Psychology and Philosophy*. 27 (New Series) (108): 389–404. JSTOR 2249035.
62. Perks, Wilfred (January 1947). "Some observations on inverse probability including a new indifference rule". *Journal of the Institute of Actuaries [JIA]*. **73**: 285–334.
63. Bayes, Thomas; communicated by Richard Price (1763). "An Essay towards solving a Problem in the Doctrine of Chances". *Philosophical Transactions of the Royal Society*. **53**: 370–418. doi:10.1098/rstl.1763.0053. JSTOR 10.2307/105741.
64. Haldane, J.B.S. (1932). "A note on inverse probability". *Mathematical Proceedings of the Cambridge Philosophical Society*. **28**: 55–61. doi:10.1017/s0305004100010495.
65. Zellner, Arnold (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley-Interscience. ISBN 978-0471169376.
66. Jeffreys, Harold (September 1946). "An Invariant Form for the Prior Probability in Estimation Problems". *Proceedings of the Royal Society*. A 24. **186** (1007): 453–461. doi:10.1098/rspa.1946.0056.
67. Berger, James; Bernardo, Jose; Sun, Dongchu (2009). "The formal definition of reference priors". *The Annals of Statistics*. **37** (2): 905–938. doi:10.1214/07-AOS587.
68. Clarke, Bertrand S.; Andrew R. Barron (1994). "Jeffreys' prior is asymptotically least favorable under entropy risk" (PDF). *Journal of Statistical Planning and Inference*. **41**: 37–60. doi:10.1016/0378-3758(94)90153-8.
69. Pearson, Karl (1892). *The Grammar of Science,*. Walter Scott, London.
70. Pearson, Karl (2009). *The Grammar of Science*. BiblioLife. ISBN 978-1110356119.

71. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman and Hall/CRC. ISBN 978-1584883883.
72. A. Jøsang. A Logic for Uncertain Probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 9(3), pp.279-311, June 2001. PDF (http://www.unik.no/people/josang/papers/Jos2001-IJUFKS.pdf)
73. H.M. de Oliveira and G.A.A. Araújo,. Compactly Supported One-cyclic Wavelets Derived from Beta Distributions. *Journal of Communication and Information Systems*. vol.20, n.3, pp.27-33, 2005.
74. Keefer, Donald L. and Verdini, William A. (1993). Better Estimation of PERT Activity Time Parameters. Management Science 39(9), p. 1086–1091.
75. Keefer, Donald L. and Bodily, Samuel E. (1983). Three-point Approximations for Continuous Random variables. Management Science 29(5), p. 595–609.
76. DRMI Newsletter, Issue 12, April 8, 2005 (http://www.nps.edu/drmi/docs/1apr05-newsletter.pdf)
77. Kruschke, John K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and Stan*. Academic Press / Elsevier. ISBN 978-0-12-405888-0.
78. Yule, G. U.; Filon, L. N. G. (1936). "Karl Pearson. 1857-1936". *Obituary Notices of Fellows of the Royal Society*. **2** (5): 72. doi:10.1098/rsbm.1936.0007. JSTOR 769130.
79. "Library and Archive catalogue". *Sackler Digital Archive*. Royal Society. Retrieved 2011-07-01.
80. "Karl Pearson sesquicentenary conference". Royal Statistical Society. 2007-03-03. Retrieved 2008-07-25.
81. David, H. A. and A.W.F. Edwards (2001). *Annotated Readings in the History of Statistics*. Springer; 1 edition. ISBN 978-0387988443.
82. Gini, Corrado (1911). "Considerazioni Sulle Probabilità Posteriori e Applicazioni al Rapporto dei Sessi Nelle Nascite Umane". *Studi Economico-Giuridici della Università de Cagliari*. Anno III (reproduced in Metron 15, 133,171, 1949): 5–41.
83. Johnson (Editor), Norman L. and Samuel Kotz (1997). *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present (Wiley Series in Probability and Statistics*. Wiley. ISBN 978-0471163817.
84. Metron journal. "Biography of Corrado Gini". Metron Journal. Archived from the original on 2012-07-16. Retrieved 2012-08-18.

External links

- "Beta Distribution" (http://demonstrations.wolfram.com/BetaDistribution/) by Fiona Maclachlan, the Wolfram Demonstrations Project, 2007.
- Beta Distribution – Overview and Example (http://www.xycoon.com/beta.htm), xycoon.com
- Beta Distribution (https://web.archive.org/web/20120829140915/http://www.brighton-webs.co.uk:80/distributions/beta.htm), brighton-webs.co.uk
- Beta Distribution Video (http://www.exstrom.com/blog/snark/posts/dancingbeta.html), exstrom.com
- Hazewinkel, Michiel, ed. (2001), "Beta-distribution", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Weisstein, Eric W. "Beta Distribution". *MathWorld*.

- Harvard University Statistics 110 Lecture 23 Beta Distribution, Prof. Joe Blitzstein (https://www.youtube.com/watch?v=UZjlBQbV1KU)