SCIMATH202
Theory of Statistics and Data Analysis

The Beta Distribution

Writing Assignment 2

Ivo Verhoeven

September 4, 2024

This writing assignment is written in partial fulfillment for the UCR course
SCIMATH202 - Theory of Statistics and Data Analysis
instructor: Dr. L.R. van den Doel
in the Spring 2019 semester.

1 Introduction

The Beta distribution, defined by the positive real parameters α and β, is a continuous probability density function defined in the interval [0,1]. Due to its existence between both upper and lower bounds, it easily lends itself to applications wherein the variable is limited to intervals themselves. Probability, as a mathematical concept, share these bounds. This motivates the interpretation of the Beta function of the probabilities of probability.

Applications of the Beta distribution are many and varied. Frequent use can be partially motivated by its innate flexibility, with special cases including the uniform, power, Gamma, F, ξ^2 and exponential distributions; for the generalised Beta distribution, this list of special cases becomes much longer [1]. For large and equal values of α and β the Beta function approximates a normal distribution [2]. The CDF of the Beta distribution is further tied to the binomial, negative binomial and Student's t distributions [3].

One application of note is its role within Bayesian statistics. Herein the Beta distribution is applied as a conjugate prior for the probabilities of Bernoulli trials [2]. This implies that after new information is made available, the Beta distributed prior results in an equally distributed posterior. The parameters are slightly modified such that $\alpha - 1$ represents the number of successes and $\beta - 1$ the number of failures in preceding trials. In the case of no prior trials, the prior collapses into the uniform distribution, thereby reflecting total ignorance (see Section 2.3). A baseball example is provided at the end of this section.

This paper aims to provide a global overview for the statistical properties of the Beta distribution. First the characterisation of the Beta distribution in terms of the Gamma and Beta functions is discussed, with a rigorous proof for the relation between these functions. The definition of the Beta distribution follows, some special cases are displayed and the CDF is provided. This is followed by a section discussing the normalisation, mean and variance. The moment generating function is provided but remains unused due to all necessary moments having already been found. Two parameter estimation techniques are discussed theoretically. The method of moments estimator is (MOME) algebraically derived. The maximum likelihood estimator is discussed and an iterative procedure for approximation is provided. Lastly these estimators are empirically verified and tested for unbiasedness, consitency and efficiency. The MLE proved less biased and was a factor of 1.22 more efficient than than the MOME, but came at a significant computational cost. For most scenarios the MOME method will prove sufficient.

Example 1.1. Batting Averages.
Adapted from [4].
It is off-season of the 2019 MLB season. Unfortunately, all the team's employment records have been destroyed in a freak accident. In attempt to identify players, management has decided to simulate an entire baseball season and assign names based on last season's batting averages (BA). Each player must go through 170 hit attempts, the expected amount of attempts for an average player per 5 seasons.

At bat is Mike Trout, an exceptionally gifted player with a batting average (BA) of .312 in the 2018 season [8]. If one considered each bat a Bernoulli trial, with a batter either hitting or missing, Trout manages to hit 31.2% of balls pitched at him. To management, however, he is just an average player and is expected to achieve a BA of only 0.248.

In terms of Bayesian statistics, assume a prior $x \sim \text{Beta}(0.248 \cdot 34 + 1, 34 - 0.248 \cdot 34 + 1)$, such that α denotes the number of successes and β denotes the failures. This distribution reflects an average baseball player and is the situation before any additional information is known. The conjugate posterior is constructed by also taking into account the hits and misses of the simulated season. The simulation depicted in Fig. 1 saw Mike Trout hit 49 times and miss 121 times. The conjugate posterior $x \sim \text{Beta}(0.248 \cdot 34 + 22, 34 - 0.248 \cdot 34 + 122)$. These distributions are displayed in Fig. 1. As expected, the distribution of Trout's BA starts moving towards his true (2018) batting average.

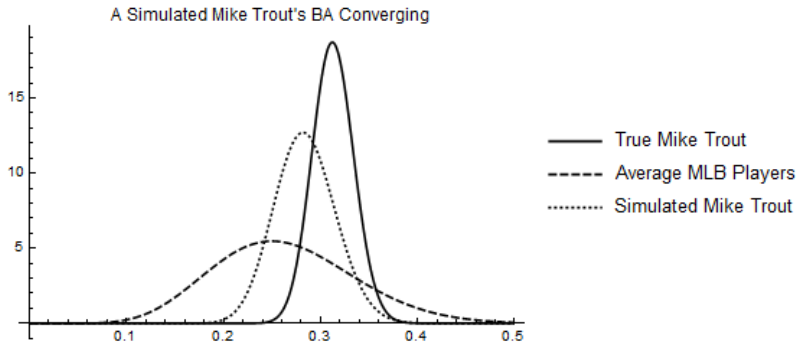

Figure 1: Mike Trout's batting average distributions before (dashed) and after (dotted) management's simulated season, with the 2018 estimate of his true BA.

2 Characterisation

2.1 The Gamma Function

Consider the discrete function of factorials such that it maps the points 1, 2, 3, 4, 5 to their factorial 1, 2, 6, 24, 120, or f[n]=n!. What is the continuous function that interpolates this discrete function?

This problem was first considered and solved by none other than Leonhard Euler, however the modern notation used is attributed to Adrien Marie Legendre [5].

$$\Gamma(x) = (n-1)! = \int_0^\infty x^{\alpha-1} e^{-x} dx \tag{1}$$

The most important property of the Gamma function is its definition through a recurrence relation. By applying integration by parts to the Gamma function as Eq. (1), the value of $\Gamma(x+1)$ can be expressed in terms of $\Gamma(x)$

$$\Gamma(x) = \int_0^\infty x^\alpha e^{-t} dt$$
$$\begin{bmatrix} f(x) = e^{-x} & \Longrightarrow & F(x) = -e^{-x} \\ G(x) = x^{x-1} & \Longrightarrow & g(x) = (x-1)x^{x-2} \end{bmatrix}$$
$$\Gamma(x) = -x^{n-1}e^{-x}|_0^\infty + (x-1)\int_0^\infty x^{x-2}e^{-x}$$

By realising that the first term tends to 0 for both the upper and lower integration bounds, and that the second term is simply the Gamma function at (x-1), the recurrence relation is found.

$$\Gamma(x+1) = (x-1)\Gamma(x-1) \implies \Gamma(x+1) = x\Gamma(x) \tag{2}$$

2.2 The Beta Function

The Beta function is very closely related to the Gamma function, in not only origin but also definition. While working on the problem described earlier, Euler first derived the Beta function (once again in the Legendre notation) [5].

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx \tag{3}$$

The Beta function can be written in terms of the Gamma function in terms of α and β as positive reals [6].

$$B(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{4}$$

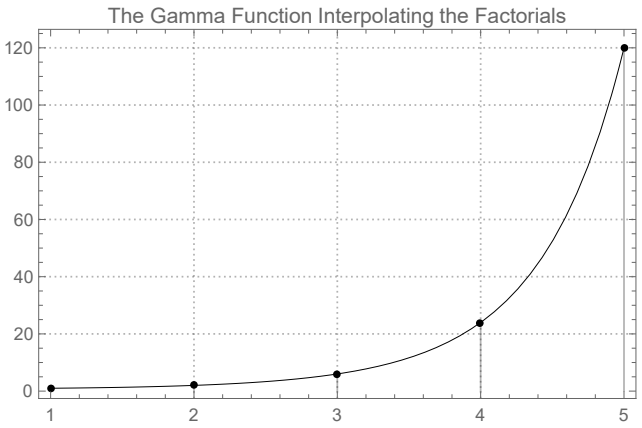

Figure 2: The Gamma function, here $\Gamma(x+1)$ to match with x, interpolating the first 5 factorials.

The proof of the relationship between the Beta and Gamma functions as per Eq. (4) follows. Consider the definition of the Gamma functions as per Eq. (1). Their product then follows as,

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty x_1^{\alpha-1}e^{-x_1}dx_1 \int_0^\infty x_1^{\beta-1}e^{-x_2}dx_1 \\ &= \int_0^\infty \int_0^\infty e^{-(x_1+x_2)}x_1^{\alpha-1}x_2^{\beta-1}dx_1dx_2 \end{aligned} \tag{5}$$

Rather than evaluate this integral, a transformation is applied. Rewrite the variable $x_1 = y_1y_2$ and $x_2 = y_1(1-y_2)$. This immediately allows one to simplify the sum of x_1 and x_2 as, $x_1+x_2 = y_1y_2+y_1(1-y_2) = y_1$. The Jacobian of this transformation follows as,

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix} = \begin{bmatrix} y_2 & y_1 \\ 1-y_2 & -x_1 \end{bmatrix} \tag{6}$$

The absolute determinant of the Jacobian follows as $|-y_2 \cdot y_1 - y_1(1-y_2)| = |-y_1| = y_1$. This allows rewriting from x_1 and x_2 into y_1 and y_2 as per the change of variables theorem [7]. This theorem states that,

$$\int\int_A f_{X_1,X_2}(x_1,x_2)dx_1dx_2 = \int\int_B f_{X_1,X_2}(w_1(y_1,y_2),w_2(y_1,y_2))|J|dy_1dy_2 \tag{7}$$

Returning to the product of two Gamma functions, it follows that in terms of x and y this becomes,

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int_{Y_2}\int_{Y_1} e^{-y_1}(y_1y_2)^{\alpha-1}(y_1(1-y_2))^{\beta-1}y_1dy_1dy_2 \\ &= \int_{Y_2} y_2^{\alpha-1}(1-y_2)^{\alpha-1}dy_2 \int_{Y_1} e^{-y_1}y_1^{\alpha-1}y_1^{\beta-1}xdx \end{aligned} \tag{8}$$

Note that the functions $\Gamma(\alpha)$ and $\Gamma(\beta)$ are both positive. This requires Y_2 to be bounded between $[0,1]$, while Y_1 can take any value between $[0,\infty]$. The integration ranges of the above integral then follow as,

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int_0^1 y_2^{\alpha-1}(1-y_2)^{\beta-1}dy_2 \int_0^\infty e^{-y_1}y_1^{\alpha-1}y_1^{\beta-1}y_1dy_1 \\ &= \int_0^1 y_2^{\alpha-1}(1-y_2)^{\beta-1}dy_2 \int_0^\infty e^{-y_1}y_1^{\alpha+\beta}dy_1 \end{aligned} \tag{9}$$

Here the left integral is equivalent to $B(u\alpha,\beta)$ defined as in Eq. (3), while the right integral is equivalent to $\Gamma(u+v)$ defined as in Eq. (1). The product of two Gamma functions therefore follows as,

$$\Gamma(\alpha)\Gamma(\beta) = B(\alpha,\beta)\Gamma(\alpha+\beta) \implies B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{10}$$

2.3 The Beta Distribution

While already applied and discussed, the derivation of the Beta distribution is presented here. This derivation shares much in common with the proof for Eq. (4), expect now using Gamma distributions rather than functions. This distribution is defined as,

$$\text{Gamma}(\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} \tag{11}$$

Consider two independent Gamma distributed random variables related as,

$$g(x_1, x_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x_1^{\alpha-1} x_2^{\alpha-1} e^{-x_1-x_2/\beta} \tag{12}$$

Rewrite the variables as $x_1 = y_1 y_2$ and $x_2 = y_1(1 - y_2)$. The Jacobian follows as,

$$J = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial y_1} \\ \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial y_1} \end{bmatrix} = \begin{bmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{bmatrix} \tag{13}$$

The absolute determinant of this Jacobian follows as $| - y_2 y_1 - y_1(1 - y_2)| = y_1$. Once again invoking the transformation of variable theorem [7], the joint PDF of y_1 and y_2 follows as,

$$\begin{aligned} g(y_1, y_2) &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} (y_1 y_2)^{\alpha-1} (y_1(1 - y_2)^{\beta-1} e^{-y_1 y_2 - \beta(y_1(1-y_2))} |J| \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha-1} y_1^{\beta-1} y_2^{\alpha-1} (1 - y_2)^{\beta-1} \end{aligned} \tag{14}$$

By integrating out the variables, the marginal PDFs can be found, considering these were independent of each other. The integration ranges follow by the same argument applied in the proof for Eq. (4), such that $y_1 \in [0, \infty]$ and that $y_2 \in [0, 1]$.

$$\int_{Y_1}^{\infty} \frac{1}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha-1} y_1^{\beta-1} dy_1 \int_0^{\infty} y_2^{\alpha-1} (1 - y_2)^{\beta-1} dy_2 \tag{15}$$

The first PDF follows as,

$$g(y_1) = \frac{1}{\Gamma(\alpha + \beta)} y_1^{\alpha+\beta-1} e^{-y_1} \tag{16}$$

Note that this is Gamma$(\alpha + \beta, 1)$, with the Gamma distribution defined as per Eq. (1). The second PDF, recalling Eq. (9), follows as,

$$\begin{aligned} g(y_2) &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \int_0^{\infty} y_1^{\alpha+\beta-1} e^{-y_1} dy_1 \cdot y_2^{\alpha-1} (1 - y_2)^{\beta-1} \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y_2^{\alpha-1} (1 - y_2)^{\beta-1} \end{aligned} \tag{17}$$

This is the function referred to by the Beta distribution. Note that with this final result, again very much like Eq. (10), one can relate the product of two Gamma distributions to the product of a Beta and a Gamma function,

$$\text{Gamma}(\alpha)\text{Gamma}(\beta) = \text{Beta}(\alpha, \beta)\text{Gamma}(\alpha + \beta) \implies \text{Beta}(\alpha, \beta) = \frac{\text{Gamma}(\alpha)\text{Gamma}(\beta)}{\text{Gamma}(\alpha + \beta)} \tag{18}$$

Finally rewriting in terms of more recognisable nomenclature, the Beta distribution is defined as,

$$\text{Beta}(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1 - x)^{\beta-1} \tag{19}$$

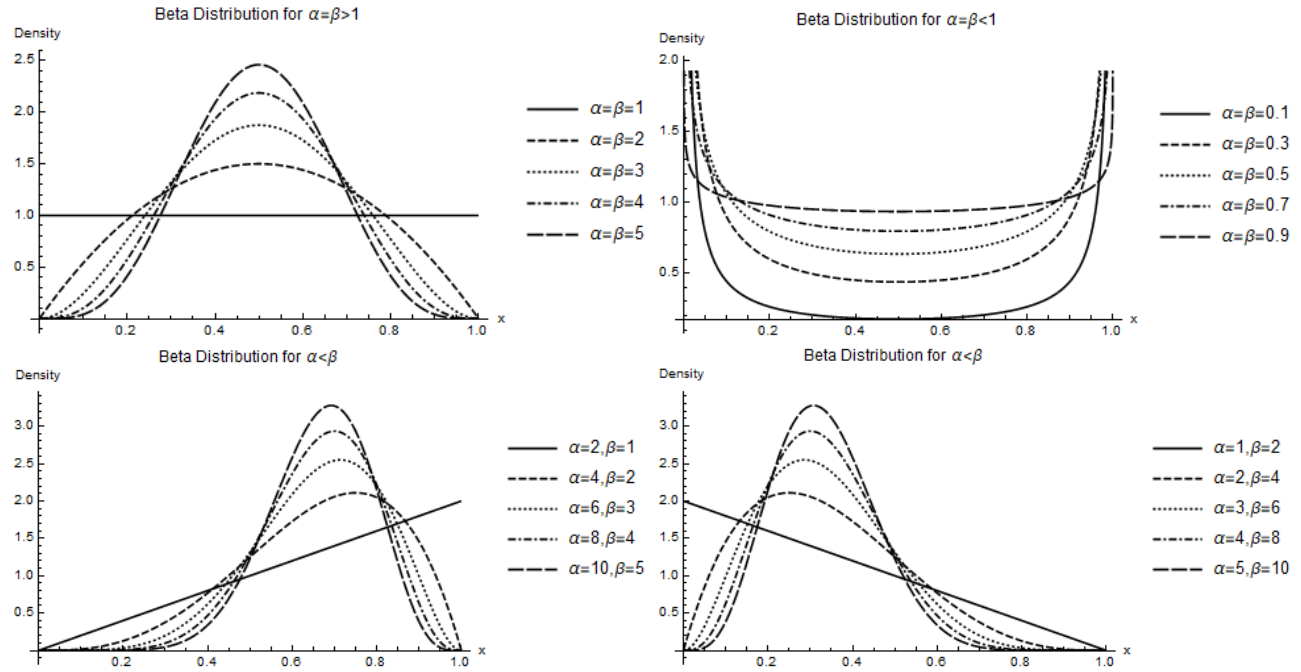

Figure 3: Four instances of the Beta distribution. For $\alpha = \beta \geq 1$ the distribution is unimodal and symmetric about $x = 0.5$. For $\alpha = \beta \leq 1$ the distribution inverts, becoming U-shaped, but remains symmetric about $x = 0.5$. For integer values of either $\alpha > \beta$ or $\alpha < \beta$ the distribution is skewed negatively or positively.

Integrating from 0 to some $x_i < 1$ gives the CDF of the Beta distribution as,

$$\begin{aligned} \textsc{Beta}(x; \alpha, \beta) &= \frac{\Gamma(\alpha+\beta)}{\mathrm{B}(\alpha,\beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{\mathrm{B}(x;\alpha,\beta)}{\mathrm{B}(\alpha,\beta)} \end{aligned} \tag{20}$$

In figure 3 a number of shapes for different cases of the Beta distribution have been displayed. For $\alpha = \beta \geq 1$ the distribution generally unimodal and centered about 0.5. As the values for α and β increase, so does it precision about its center. A special case is for $\alpha = \beta = 1$: this is the uniform distribution for $0 \leq 1$. Generally for equal values for α and β the cumulative probabilities can be estimated by the normal distribution [2]. For $\alpha = \beta < 1$ U-shaped, with extrema at both domain boundaries, but remains centered about 0.5. Another special case is for $\alpha = \beta = 0.5$: this is the arcsine distribution. For unequal values of α and β the distribution becomes skewed towards the more dominant parameter. The two corresponding plots in figure 3 clearly show the symmetric nature of the Beta distribution.

3 Statistical Properties

3.1 Normalisation

From the definition of the cumulative Beta function as per Eq.(20) the proof of normalisation becomes trivial.

$$\frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{1}{B(\alpha,\beta)} \cdot B(\alpha,\beta) \equiv 1 \tag{21}$$

This makes sense given the intuitive definition of the (cumulative) Beta distribution. Computing the probability of some $x_i \sim \text{Beta}(\alpha, \beta)$ is simply computing the Beta function for the integration range of 0 to some $x_i < 1$ divided by the Beta function computed over the whole possible domain.

3.2 Mean

Let the mean of x, μ_x be defined by the first expectation value. Taking X to follow a Beta distribution, the mean of X can be calculated as follows.

$$\begin{aligned}\mathrm{E}\{x\} &= \int_X x \cdot f(x)dx = \int_0^1 x \cdot \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1}dx \end{aligned} \tag{22}$$

Note that the above integral closely approximates Eq. (3) with $\alpha+1$ as its first parameter. This allows for rewriting as

$$\mathrm{E}\{x\} = \frac{1}{B(\alpha,\beta)} B(\alpha+1,\beta)$$

Recall the relation of the Gamma function to the Beta function, as per Eq. (4). Rewriting in terms of the Gamma function rather than the Beta function provides a result that is more easily simplified.

$$\begin{aligned}\mathrm{E}\{x\} &= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta+1)/(\Gamma(\alpha+1)\Gamma(\beta))}{\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))} \\ &= \frac{\Gamma(\alpha+1)\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+1)}\end{aligned}$$

The Gamma function can be defined using the following recurrence relationship as per Eq. (2).

$$\Gamma(x+1) = x\Gamma(x)$$

The mean of X is then easily found by rewriting in terms of $\Gamma(\alpha)$ and $\Gamma(\beta)$ and simplifying.

$$\begin{aligned}\mathrm{E}\{x\} &= \frac{\Gamma(\alpha+1)\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+1)} \\ &= \frac{\alpha\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)(\alpha+\beta)\Gamma(\alpha+\beta)} \\ \mu_x &= \frac{\alpha}{\alpha+\beta}\end{aligned} \tag{23}$$

Given that both α and β are strictly positive reals, it follows that for all values of these parameters the mean of x will fall within the [0,1] domain.

3.3 Variance

Let the variance of x, $var\{x\}$, be the difference of the expectation value of X^2 and the squared mean. Due to the repetitive nature of this derivation, some steps explained in the derivation of $\mathrm{E}\{x\}$ while be applied but not shown or highlighted.

$$\begin{aligned}\mathrm{E}\{x^2\} &= \frac{1}{B(\alpha,\beta)} \int_0^1 x^2 x^{\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha+1}(1-x)^{\beta-1}dx \\ &= \frac{B(\alpha+2,\beta)}{B(\alpha,\beta)} \\ &= \frac{(\alpha+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)(\alpha+\beta+1)\Gamma(\alpha+\beta)} \\ \mathrm{E}\{x^2\} &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\end{aligned} \tag{24}$$

The variance then follows as the difference of Eq. (24) and the square of Eq.(23).

$$\begin{aligned}\mathrm{var}\{x\} &= \mathrm{E}\{x^2\} - \mathrm{E}\{x\}^2 = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\ &= \frac{\alpha(\alpha+1)(\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)} - \frac{\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{\alpha^3+\alpha^2\beta+\alpha^2+\alpha\beta-\alpha^3-\alpha^2\beta-\alpha^2}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ \mathrm{var}\{x\} &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned} \tag{25}$$

3.4 Moment Generating Function

The moment generating function (MGF) is defined as the expectation value of the Laplace transform of the probability density function. One important use is the calculation of the PDF's moments, through computing the n-th derivative at $t = 0$ for the n-th moment.

$$\begin{aligned} m_X(t) = \mathrm{E}\{e^{tx}\} &= \int_0^1 e^{tx} \cdot f(x;\alpha,\beta) \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 e^{tx} \cdot x^{\alpha-1}(1-x)^{\beta-1} \end{aligned} \tag{26}$$

A form independent of can be found by using the Taylor series approximation of e^{tx} and recalling the definition of the Beta function as per Eq. (3).

$$\begin{aligned} m_X(t) &= \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 (\sum_{k=0}^{\infty} \frac{t^k x^k}{k!}) \cdot x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)} \sum_{k=0}^{\infty} \frac{t^k}{k!} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1} \\ &= \sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{\mathrm{B}(\alpha+k,\beta)}{\mathrm{B}(\alpha,\beta)} \\ &= 1 + \sum_{k=1}^{\infty} \frac{t^k}{k!} \frac{\mathrm{B}(\alpha+k,\beta)}{\mathrm{B}(\alpha,\beta)} \end{aligned} \tag{27}$$

The quotient of these Beta functions can be solved when inserting the Gamma functions as per Eq. (4). The final result is the confluent hypergeometric function of the first kind [9].

$$m_X(t) = 1 + \sum_{k=1}^{\infty} \frac{t^k}{k!} \prod_{n=0}^{k-1} \frac{\alpha+n}{\alpha+\beta+n} \tag{28}$$

4 Estimation of Parameters

4.1 Method of Moments Estimation

The method of moments requires setting the moments of the PDF equal to the moments derived from a data set of length n, $x_1, x_2, \ldots, x_n$. For the first two moments, the mean and variance as given by Eq. (23) and Eq. (25) are thus set to the sample mean and variance,

$$\mu_x = \overline{x}, \mathrm{var}\{x\} = s_x^2 \implies \overline{x} = \frac{\alpha}{\alpha+\beta}, s_x^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{29}$$

These relations can then be reformulated to describe the parameters α and β solely in terms of the sample mean and variance. For the sample mean, this follows as,

$$\begin{aligned}\overline{x} &= \frac{\alpha}{\alpha+\beta} \\ &\implies (\alpha+\beta)\overline{x} = \alpha \\ &\implies \alpha - \alpha\overline{x} = \beta\overline{x} \\ &\implies \beta = \frac{\alpha}{\overline{x}} - \alpha\end{aligned}$$

Attempting the same for the sample variance will yield an expression independent of β for α.

$$\begin{aligned}s^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &\implies \alpha\beta = (\alpha+\beta)^2(\alpha+\beta+1)s^2 \\ &\implies \alpha(\frac{\alpha}{\overline{x}} - \alpha) = (\alpha + \frac{\alpha}{\overline{x}} - \alpha)^2(\alpha + \frac{\alpha}{\overline{x}} - \alpha + 1)s^2 \\ &\implies \alpha^2\frac{1}{\overline{x}} - \alpha^2 = \frac{\alpha^2}{\overline{x}^2}(\frac{\alpha}{\overline{x}} + 1)s^2 \\ &\implies \frac{1}{\overline{x}} - 1 = \frac{1}{\overline{x}^2}(\frac{\alpha}{\overline{x}} + 1)s^2 \\ &\implies (\frac{1-\overline{x}}{\overline{x}})\frac{\overline{x}}{s^2} = \frac{\alpha}{\overline{x}} + 1 \\ \hat{\alpha} &= \overline{x}(\frac{\overline{x}(1-\overline{x})}{s^2} - 1)\end{aligned} \tag{30}$$

All that remains is expressing β using the result for α,

$$\begin{aligned}\beta &= \frac{\alpha}{\overline{x}} - \alpha \\ &\implies \beta = \frac{\alpha(1-\overline{x})}{\overline{x}} \\ &\implies \beta = \overline{x}(\frac{\overline{x}(1-\overline{x})}{s^2} - 1)\frac{(1-\overline{x})}{\overline{x}} \\ \hat{\beta} &= (1-\overline{x})(\frac{\overline{x}(1-\overline{x})}{s^2} - 1)\end{aligned} \tag{31}$$

This estimator is unbiased. Assuming the sample mean and variance to be unbiased -E$\{\overline{x}\} = \mu_x$ and E$\{\overline{s^2}\}$ = var$\{x\}$- unbiaseness can be proven by inputting Eq. (23) and Eq. (25) into Eq. (30) and Eq. (31).

4.2 Maximum Likelihood Estimation

The estimation method employing maximum likelihood (MLE) requires the calculation of the likelihood of parameters based on an i.i.d. sampled data set from the target distribution. The likelihood function is defined as the product of the probabilities of each value within the data set given the distribution in terms of its unknown parameters. Rather than maximising this function, its log is taken, often making subsequent calculations much simpler. For the Beta distribution the likelihood function follow as,

$$\begin{aligned}L(\alpha,\beta) &= \prod_{i=1}^{n}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x_i^{\alpha-1}(1-x_i)^{\beta-1} \\ &= (\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)})^n\prod_{i=1}^{n}x_i^{\alpha-1}(1-x_i)^{\beta-1}\end{aligned} \tag{32}$$

Again, it is often much simpler to work with the log-transform of the likelihood function, given that derivatives are required when maximizing. For the Beta distribution this follows as,

$$\begin{aligned} l(\alpha,\beta) &= \log(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)})^n \prod_{i=1}^{n} x_i^{\alpha-1}(1-x_i)^{\beta-1} \\ &= n\log(\Gamma(\alpha+\beta)) - n\log(\Gamma(\alpha)) - n\log(\Gamma(\beta)) \\ &\quad + (\alpha-1)\sum_{i=1}^{n}\log(x_i) + (\beta-1)\sum_{i=1}^{n}\log(1-x_i) \end{aligned} \tag{33}$$

To maximise the likelihood, the partial derivatives in terms of α and β are set to 0 and solved for $\hat{\alpha}$ and $\hat{\beta}$. The functions that need to be solved are displayed below. Immediately the interdependence of this system become obvious: one can't solve for α without β, and vice versa. A closed form solution does not exist, however an iterative method can be applied to approximate the values for α and β [10].

$$\frac{\partial}{\partial\alpha}l(\alpha,\beta) = n(\frac{\Gamma^{(1)}(\alpha+\beta)}{\Gamma(\alpha+\beta)}) - n(\frac{\Gamma^{(1)}(\alpha)}{\Gamma(\alpha)}) + \sum_{i=1}^{n}\log(x_i) \tag{34}$$

$$\frac{\partial}{\partial\beta}l(\alpha,\beta) = n(\frac{\Gamma^{(1)}(\alpha+\beta)}{\Gamma(\alpha+\beta} - n(\frac{\Gamma^{(1)}(\beta)}{\Gamma(\beta)}) + \sum_{i=1}^{n}\log(1-x_i) \tag{35}$$

The digamma function is defined as the derivative of the natural logarithm of the Gamma function. This makes notation much easier.

$$\psi(x) = \log(\Gamma(x)) = \frac{\partial}{\partial x}\frac{\Gamma^{(1)}(x)}{\Gamma(x)} \tag{36}$$

Owen [10] recommends applying the two-dimensional Newton-Raphson method to root finding of the system defined by Eq. (34) and Eq. (35). Let $\vec{g}$ be the vector with these equations, rewritten in terms of the digamma function, and $\vec{G}$ be the Hessian matrix (similar to the Jacobian but with second derivatives).

$$\vec{g} = \begin{bmatrix} \psi(\alpha) - \psi(\alpha+\beta) - \sum_{i=1}^{n}\log(x_i) \\ \psi(\beta) - \psi(\alpha+\beta) - \sum_{i=1}^{n}\log(1-x_i) \end{bmatrix} \tag{37}$$

$$\vec{G} = \begin{bmatrix} \frac{\partial g1}{\partial\alpha} & \frac{\partial g1}{\partial\beta} \\ \frac{\partial g2}{\partial\alpha} & \frac{\partial g2}{\partial\beta} \end{bmatrix} = \begin{bmatrix} \psi^{(1)}(\alpha) - \psi^{(1)}(\alpha+\beta) & -\psi^{(1)}(\alpha+\beta) \\ -\psi^{(1)}(\alpha+\beta) & \psi^{(1)}(\beta) - \psi^{(1)}(\alpha+\beta) \end{bmatrix} \tag{38}$$

The parameters can then be approximated iterative by subtracting the product of the inverse Hessian matrix ($\vec{G}^{-1}$) and the system of differential equations ($\vec{g}$) from the current estimate. As $i \to \infty$, the estimate converges to $\hat{\alpha}_M LE$ and $\hat{\beta}_M LE$. In simulations 100 iterations were used, although 10 proved sufficient for most cases.

$$\{\hat{\alpha},\hat{\beta}\}_{i+1} = \{\hat{\alpha},\hat{\beta}\}_i - \vec{G}^{-1}\cdot\vec{g} \tag{39}$$

4.3 Efficiency

Thr Cramer-Rao Lower Bound (CRLB) states that the variance of an estimator is always above or at the inverse of the Fisher information of the parameters. Under the assumption that θ is an unbiased estimator of the true parameter, it follows as

$$\mathrm{var}\{\hat{\theta}\} \geq \frac{1}{I_X(\theta)} \tag{40}$$

Here $I_X(\theta)$ is the Fisher information function in terms of θ, defined as,

$$I_X(\theta) = -\mathrm{E}\{\frac{\partial^2}{\partial^2\theta}\log l(x;\theta)\} \tag{41}$$

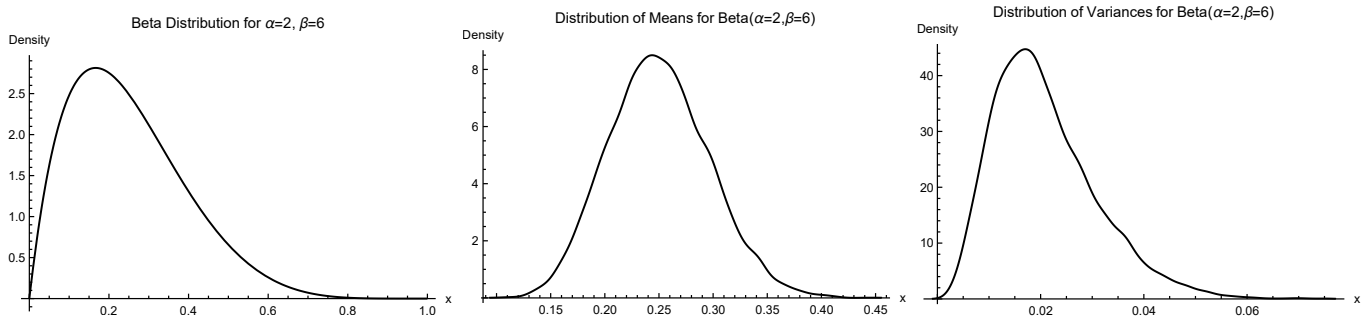

Figure 4: Distribution of Beta(2,6) and the distribution of its means and variances for 10 000x10 samples.

The log-likelihood functions are given by Eq. (34) and Eq. (35). Their derivatives, in terms of α and β, are independent in terms of x_i. These are given by the diagonal elements of (38).

$$I(\alpha) = \psi^{(1)}(\alpha) - \psi^{(1)}(\alpha + \beta) \tag{42}$$

$$I(\beta) = \psi^{(1)}(\beta) - \psi^{(1)}(\alpha + \beta) \tag{43}$$

5 Simulation of Estimators

5.1 Procedure

In order to simulate the properties of the method of moments (MOME) and maximum likelihood (MLE) estimators, a Beta distribution was sampled 1000 times with parameters $\alpha = 2$ and $\beta = 6$. The distribution of Beta(2,6) and the sample means and variance have been displayed in fig. 4. For the behaviour of these estimators under sample size, sample size from the interval [25,2500] were drawn at each multiple of $n = 25$. Each estimator iterated 100 times at different samples to calculate the expected variance. For estimating the bias, the mean of estimates was subtracted from the true value of the estimate. For estimating the efficiency of the estimators, the mean ratio of the variance over the CRLB at each sample size was taken.

$$\text{Bias} = \theta - \hat{\theta} \tag{44}$$

$$\text{Efficiency} = \text{mean}\{\frac{\text{Var}\{\hat{\theta}\}}{\text{CRLB}}\} \tag{45}$$

$$\tag{46}$$

5.2 Results

For both the MOME and MLE methods, the results appear to be symmetrically distributed about the true parameter and steadily decreasing with sample size. For both cases, the variance of the estimate for β is larger, however this was to be expected given that the CRLB was positively related to the parameter values and that $\beta > \alpha$ for these simulations. From figures 5(c,d) and 5(c,d) the estimators appear consistent, with variance decrease as a function of sample size.

Table 1: Properties of Simulated Results

		MOME		MLE	
Parameters		Bias	RE	Bias	RE
α	2	-3.01×10^{-3}	4.49	-0.40×10^{-3}	3.66
β	6	-9.95 ×10^{-3}	4.2	-5.62×10^{-3}	3.44

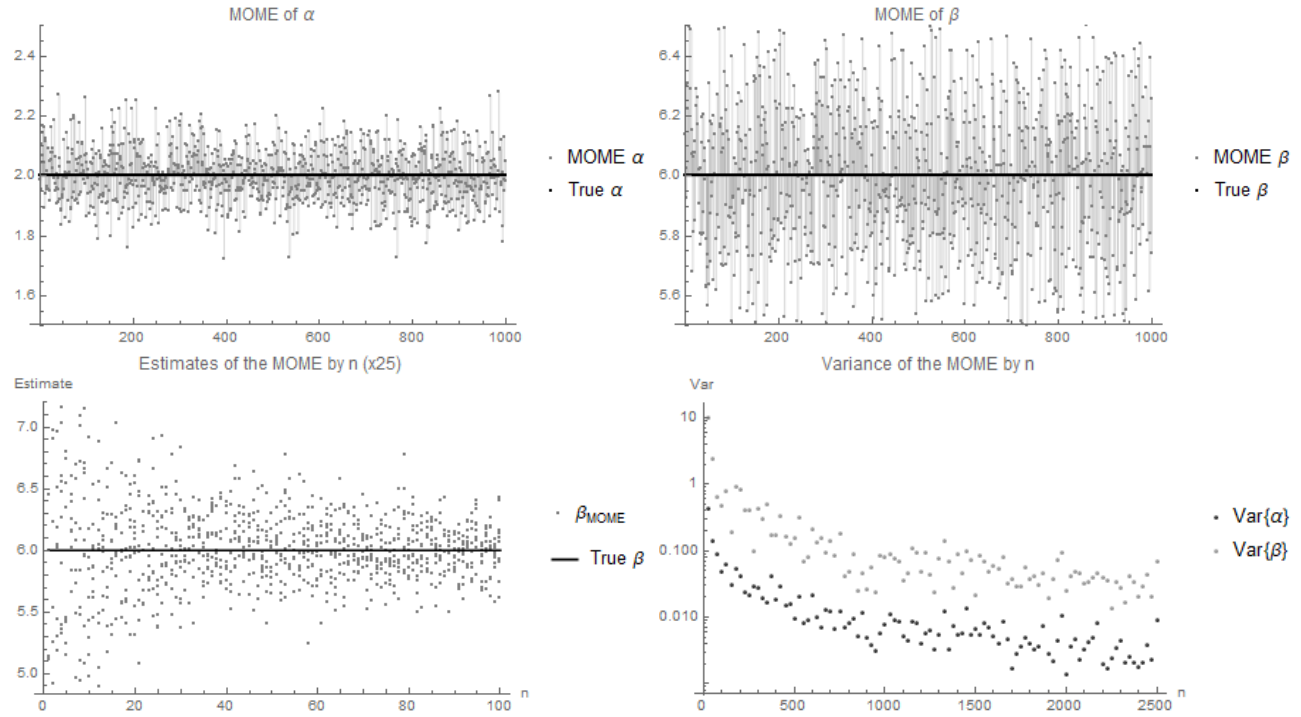

Figure 5: Method of Moments Estimation of the parameters $\alpha = 2$ and $\beta = 6$. First many estimates are shown. Then the variance of the estimate is shown for different values of sample size.

The bias for both estimator methods was negligible, however the MLE method achieved lower values for both parameters. Neither estimator was efficient, being a mean factor of 4 and 3 greater than the CRLB (rounded). The MLE method did prove more relatively efficient, being a factor of 1.22 closer to the CLRB than the MOME method. This came at a great computational cost however, requiring a computer for any reasonable sample size. With the difference between the MLE and MOME being so small, in both bias and efficiency, the MOME method will prove sufficient for virtually all applications.

6 Conclusion

This paper has provided a comprehensive overview and derivation of the statistical properties of the Beta distribution. The most important of these included the display of special cases, the derivation of the mean and the derivation of the variance. After a purely theoretical discussion of the underpinnings of this distribution, two estimators were put into practise. While the method of moments estimator was easy to derive, it was outperformed by the maximum likelihood estimator. It must be noted however, that the MLE method was purely analytical and requires the use of a computer for application to a data set of any meaningful size. For data sets above $n = 1000$ Wolfram Mathematica struggled with generating output.

Regardless, both estimators proved to be unbiased, consistent but inefficient. The slight edge of the MLE over the MOME method is in line with previous literature, however it was suggested that due to iterative estimation, the parameters $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$ are incredibly sensitive to starting conditions [10]. This was also the case in the simulations performed for this paper. Therefore, it is recommended to use the MOME for estimating the parameters of the Beta distribution.

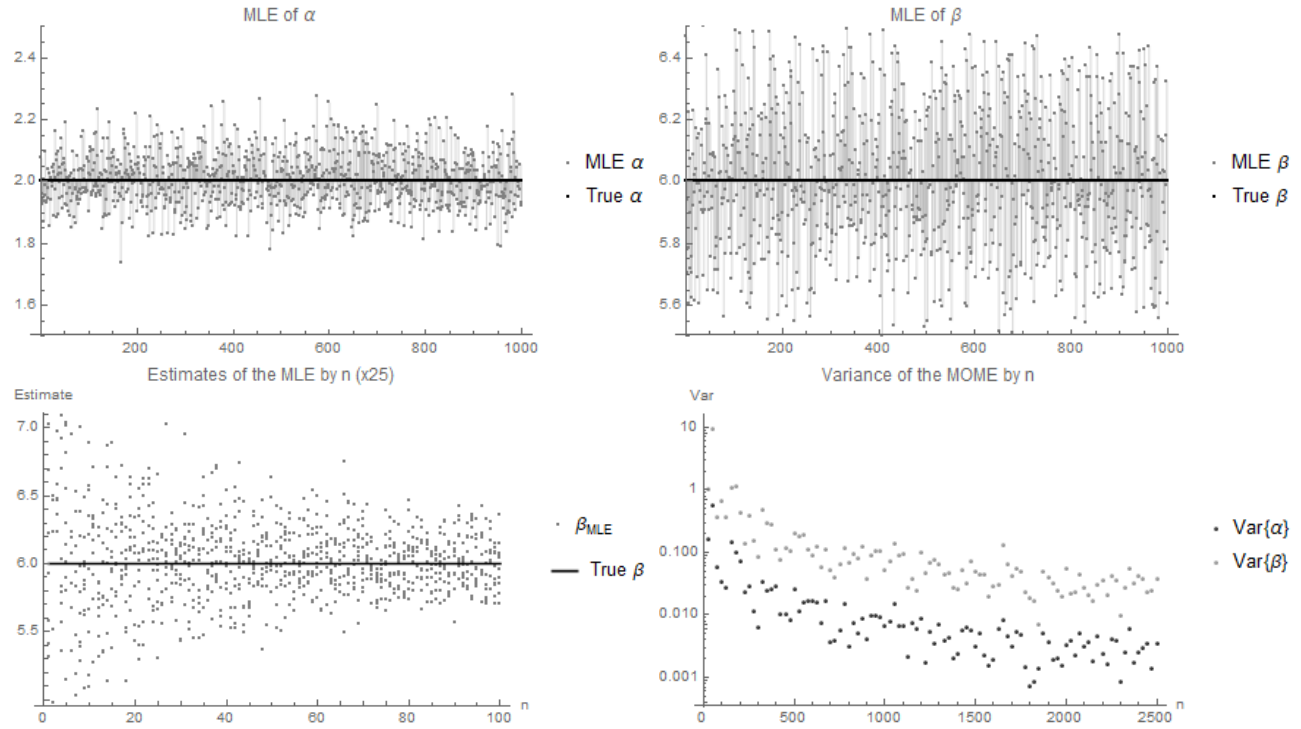

Figure 6: Maximum Likelihood Estimation of the parameters $\alpha = 2$ and $\beta = 6$. First many estimates are shown. Then the variance of the estimate is shown for different values of sample size.

References

[1] McDonald, J. B., Xu, Y. J. (1995). A generalization of the beta distribution with applications. Journal of Econometrics, 66(1-2), 133-152.

[2] Krishnamoorthy, K. (2016). *Handbook of statistical distributions with applications*. Chapman and Hall/CRC.

[3] Walck, C. (1996). Hand-book on statistical distributions for experimentalists (No. SUF-PFY/9601).

[4] Robinson, D. (2014). Understanding the beta distribution (using baseball statistics). Retrieved from: http://varianceexplained.org/statistics/betadistributionandbaseball.com

[5] Davis, P. J. (1959). Leonhard Euler's integral: A historical profile of the Gamma function *The American Mathematical Monthly*, 66(10), 849-869.

[6] Weisstein, Eric W. "Beta Function." From MathWorld–A Wolfram Web Resource. Retrieved from: http://mathworld.wolfram.com/BetaFunction.html

[7] Hogg, R. V., McKean, J., Craig, A. T. (2013). *Introduction to mathematical statistics*. Pearson Education.

[8] Mike Trout Stats. (2019). Retrieved from: https://www.baseball-reference.com/players/t/troutmi01.shtml

[9] Taboga, M. (nd.). Beta distribution. Retrieved from: https://www.statlect.com/probability-distributions/beta-distributionhid6

[10] Owen, C. E. B. (2008). Parameter estimation for the beta distribution. MSc Thesis, Department of Statistics, Brigham Young University.