Lab 7/8 – MATH 240 – Computational Statistics

Jake Schneider
Colgate University
Mathematics
jdschneider@colgate.edu

Abstract

This lab report focuses on understanding and applying the beta distribution. We examined its theoretical properties by computing the mean, variance, skewness and kurtosis across four parameter settings. Using both numerical integration and simulated data, we verified the population moments and explored how sample statistics converge to their theoretical values. We applied these concepts to real world data by modeling country level death rates. We estimated the distribution parameters using both Method of Moments (MOM) and Maximum Likelihood Estimation (MLE). Overall this lab highlights the flexibility of the beta distribution and its usefulness in modeling bounded, proportion based data.

Keywords: Beta Distribution, Point Estimation

1 Introduction to Beta Distribution

The beta distribution is a flexible, continuous probability distribution used to model a random variable X on the interval [0,1] or (0,1) in terms of two positive parameters, alpha and beta. These two parameters control the shape of the distribution. The properties of the beta distribution make it useful for modeling proportions, probabilities or rates due to its ability to take on a wide range of forms.

2 Density Functions and Parameters

The beta distribution's probability density function is defined as:

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I(x \in [0,1])$$

In our function α and β are the shape parameters and Γ is the gamma function. This allows the beta distribution to be remarkably flexible with regards to its shape; it can be left-skewed, right skewed or symmetric depending on the value of the parameters that define its shape which we can see if we look to Figure 1 which illustrates how the shape can change across the four different parameter combinations. Summary statistics for each of these cases are presented in Table 3, offering a clearer understanding of how the distribution behaves under various conditions, which we also summarize below.

- α>1, β>1: The distribution is uni modal and resembles a bell curve.
- α<1, β<1: The distribution is U-shaped, with a higher density near 0 and 1.
- α=β=1: The distribution simplifies to a uniform distribution over [0,1].
- α>1, β<1: The distribution skews towards 1.
- α<1, β>1: The distribution skews towards 0.

3 Properties

These are several of the key properties for the beta distribution.

$$\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta} \qquad \text{(Mean)}$$

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \text{(Variance)}$$

$$\text{Skew}(X) = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}} \qquad \text{(Skewness)}$$

$$\text{Kurt}(X) = 6 \cdot \frac{(\alpha-\beta)^2(\alpha+\beta+1) - \alpha\beta(\alpha+\beta+2)}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)} \qquad \text{(Excess Kurtosis)}$$

These formulas are used to compute the true values of the mean, variance, skewness, and excess kurtosis for our various parameter combinations. To verify them numerically, we created a R function called `beta.moment()`, that computes both centered and uncentered arguments using numerical integration. We can see compare our numerical summaries in Table 3 to our theoretical summaries in Table 2.

In addition, by generating a large number of samples from a know beta distribution with parameters α=2 and β=5 we demonstrated that the sample based estimates of these statistics being to converge to their true population-level statistics which we can see in Figure 2, which was done using the `cumstats` package (Erdely and Castillo, 2017). This convergence demonstrates the Law of Large Numbers. Furthermore, when we modeled the sampling distribution of these statistics we see that in Figure 3 that our resulting distributions are narrow and symmetric. This would indicate to us that our estimates were not only accurate but also precise. Together, these results support the reliability of using sample based summaries to approximate population characteristics for the Beta Distribution.

4 Estimators

When applying the Beta distribution, estimating the parameters α and β is critical for real world application. The two methods that we can use for estimating these parameters are:

- Method of Moments (MOM): This method equates the first k sample moments to the first k population moments of the Beta Distribution to solve for α and β.
- Maximum Likelihood Estimation (MLE): This approach creates a likelihood function based on the distributions PDF and finds that parameter value that maximizes the probability of observing the data.

To assess the performance of our estimators we can compare the estimates using bias, precision and mean squared error (MSE). Bias measures the average deviation from the true parameter value, precision reflects the consistency of the estimates, and MSE captures the overall error by combining both bias and variance. These three metrics provide insight into how well each method performs in estimating the underlying parameters of the distribution. Given a vector of estimates `theta.hats` where the true values is `theta`, which we can compute as follows.

$$\text{Bias} = \mathbb{E}(\hat{\theta}) - \theta$$

$$\text{Precision} = \frac{1}{\text{Var}(\hat{\theta})}$$

$$\text{MSE} = \text{Var}(\hat{\theta}) + \Big(\mathbb{E}(\hat{\theta}) - \theta\Big)^2$$

5 Example: Modeling Death Rates

To apply the beta distribution to a real world setting, we collected data from the World Bank on country's death rates in 2022. Since death rates are naturally bounded between 0 and 1 when being expressed as proportions the beta distribution would be a good distribution for this data.

5.1 Collecting Data and Estimating Parameters

The data we collected reported death rates per 1000 citizen, and the support for the beta distribution is (0,1), so we need to convert this rate into a proportion to ensure they fall are within the support. To clean the data we selected the country name, code and year 2022. Once the data was collected and cleaned we then estimated the parameters α and β using the both the MOM and MLE techniques. For the MOM technique used the `nleqslv` package (Hasselman, 2023) and use the function `nleqslv()`. As for the MLE method we use the `optim()` function to maximize the log-likelihood function.

To evaluate the the performance of each method, we simulated 1000 iterations using a `for()` loop with α=8 and β=950 and used `set.seed(7272+i)`. This estimated 1000 method of moments estimates for α and β, and maximum likelihood estimates for α and β. We can see the distribution of our estimates if we look to Figure 4. We also created table to summarize the accuracy and precision of our estimates to help us determine which method we should use in this case which we can see in Table 1.

5.2 Results

The simulation results collected show a distinct difference between the MOM and MLE estimators. The distribution of MLE estimators for both α and β are more concentrated around the true parameter values compared to the MOM estimates. This is visible in our density plots in Figure **??** which shows a more narrow and symmetric curve for both parameters of the MLE function. The table below also shows a numerical summary of the performance for each of the techniques.

Table 1: Summary Statistics Estimator Distribution

	Method	Parameter	Bias	Precision	MSE
1	MOM	Alpha	0.08	1.83	0.55
2	MOM	Beta	10.57	0.00	8303.00
3	MLE	Alpha	0.07	2.13	0.47
4	MLE	Beta	9.18	0.00	7118.73

We see in our table that MLE outperformed MOM with a lower bias, higher precision and smaller MSE for all of the parameters, that is we would opt to use the MLE technique as the most accurate method to estimate our parameters.

5.3 Discussion

This examples demonstrates how the beta distribution can be applied to real world examples of proportions that are naturally bounded between 0 and 1. Through both theoretical and applied analysis, we gained insight into the flexibility and the behavior given different parameters.

Both MOM and MLE offered estimates that would we could have used for our distributions parameters. See saw that MOM is simpler to implement and interpret its estimates, were more variable and less accurate compared to MLE. Now that we have successful fitted a beta distribution to real data, we can now start to as some questions about the data. For example we may ask, what is the probability that a country's death rate exceeds a certain percentage? Or, if the World Health Organization wanted to allocate more resources to the worst 10% of countries what would the death rate cutoff be? While these are not questions that we will answer in this lab it is important to consider the important application that our information may be able to provide us to help answer real world problems. To answer these question to the best of our ability it is essential that we establish a strong statistical foundation to provide the most accurate results.

References

Erdely, A. and Castillo, I. (2017). *cumstats: Cumulative Descriptive Statistics.* R package version 1.0.

Hasselman, B. (2023). *nleqslv: Solve Systems of Nonlinear Equations.* R package version 3.3.5.

6 Appendix

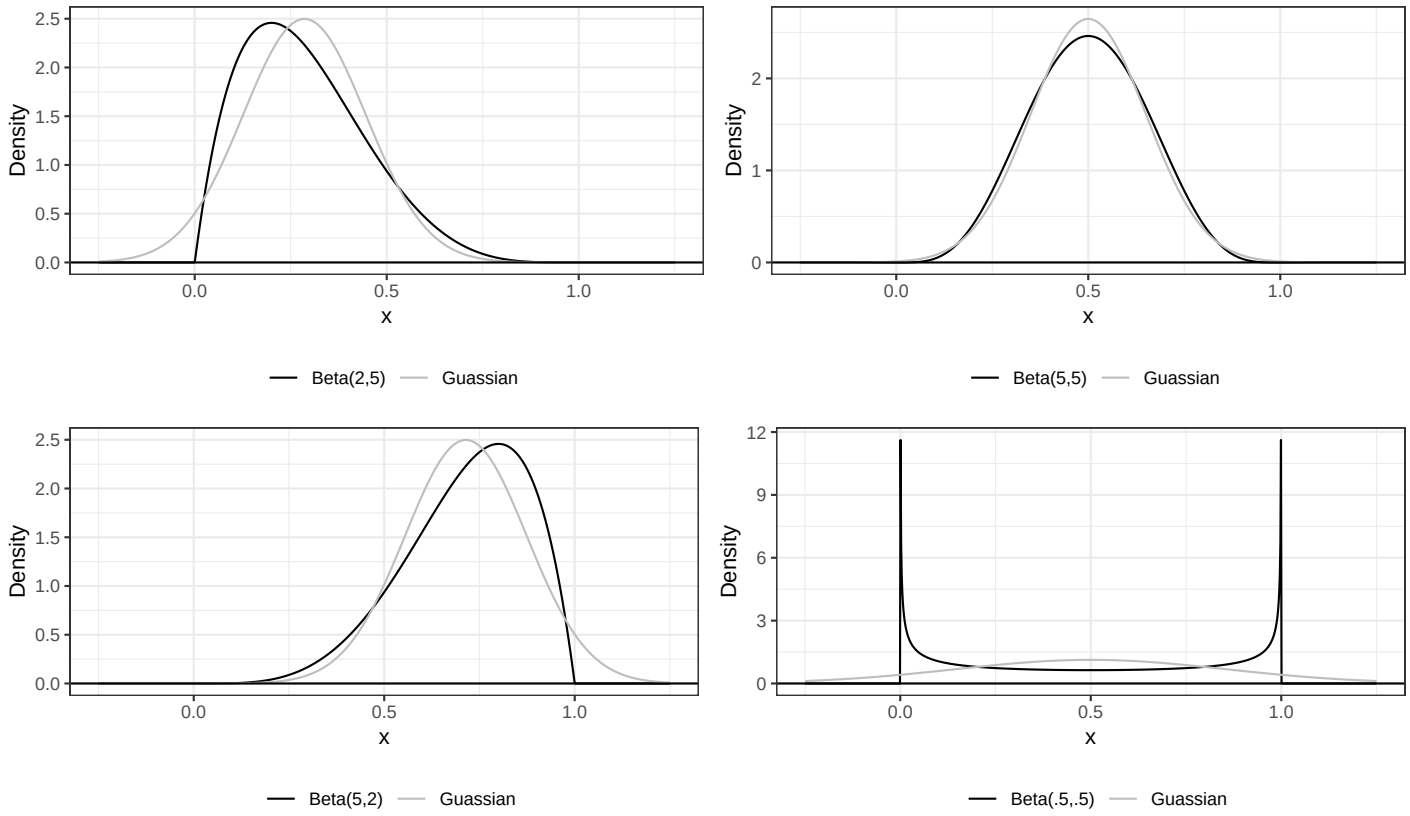

Figure 1: 4 Beta Distribution Cases

Table 2: Population Level Summary Statistics Of 4 Beta Distribution Parameters

	Alpha	Beta	Mean	Variance	Skewness	Excess Kurtosis
1	2.00	5.00	0.29	0.03	0.60	-0.12
2	5.00	5.00	0.50	0.02	0.00	-0.46
3	5.00	2.00	0.71	0.03	-0.60	-0.12
4	0.50	0.50	0.50	0.12	0.00	-1.50

Table 3: Numeical Summary Statistics of 4 Beta Dsitribution Parameters

	Alpha	Beta	Sample Mean	Sample Variance	Sample Skewness	Sample Excess Kurtosis
1	2.00	5.00	0.29	0.03	0.57	-0.23
2	5.00	5.00	0.50	0.02	-0.06	-0.58
3	5.00	2.00	0.71	0.03	-0.57	-0.23
4	0.50	0.50	0.51	0.13	-0.02	-1.51

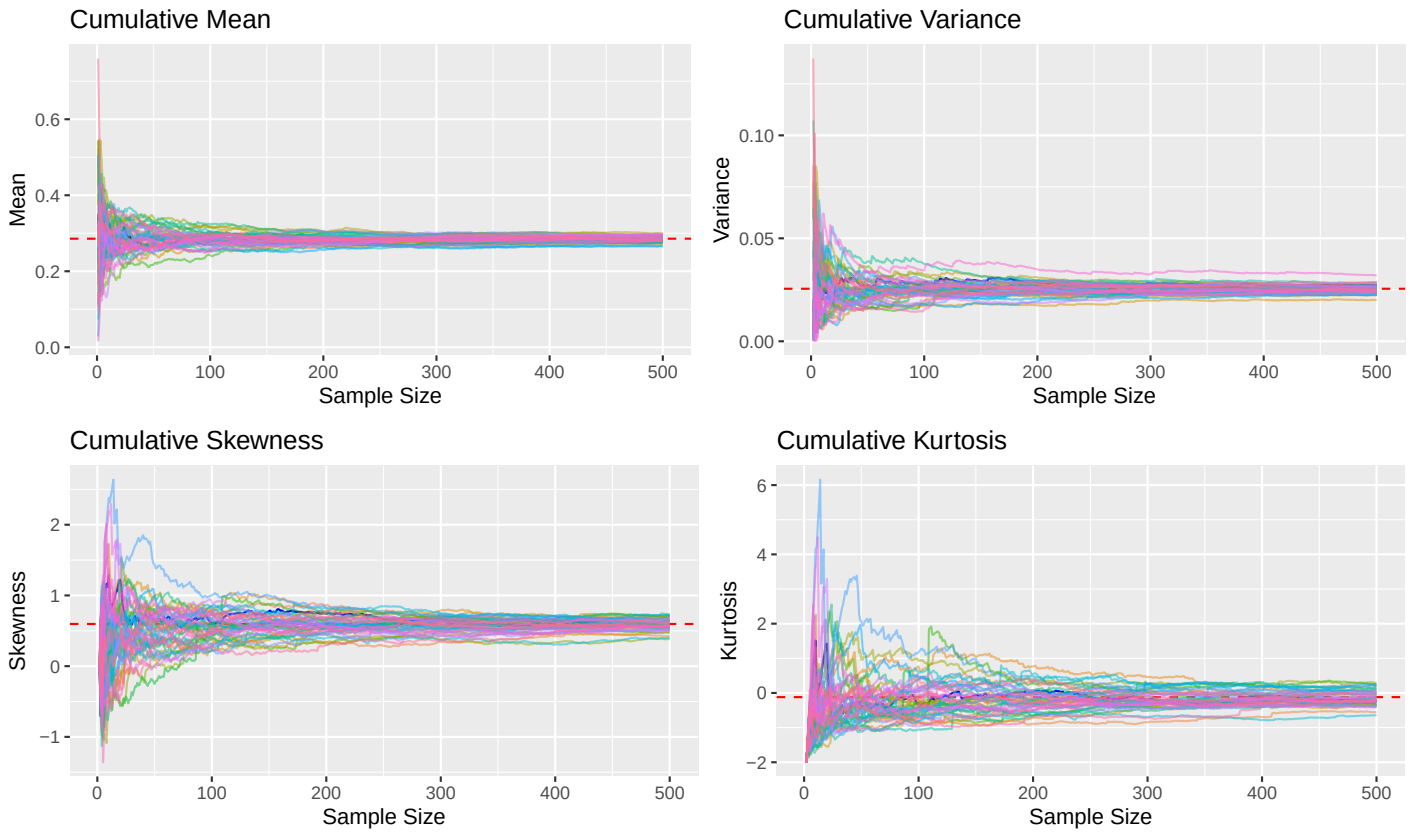

Figure 2: α=2 and β=5 Convergence Simulation

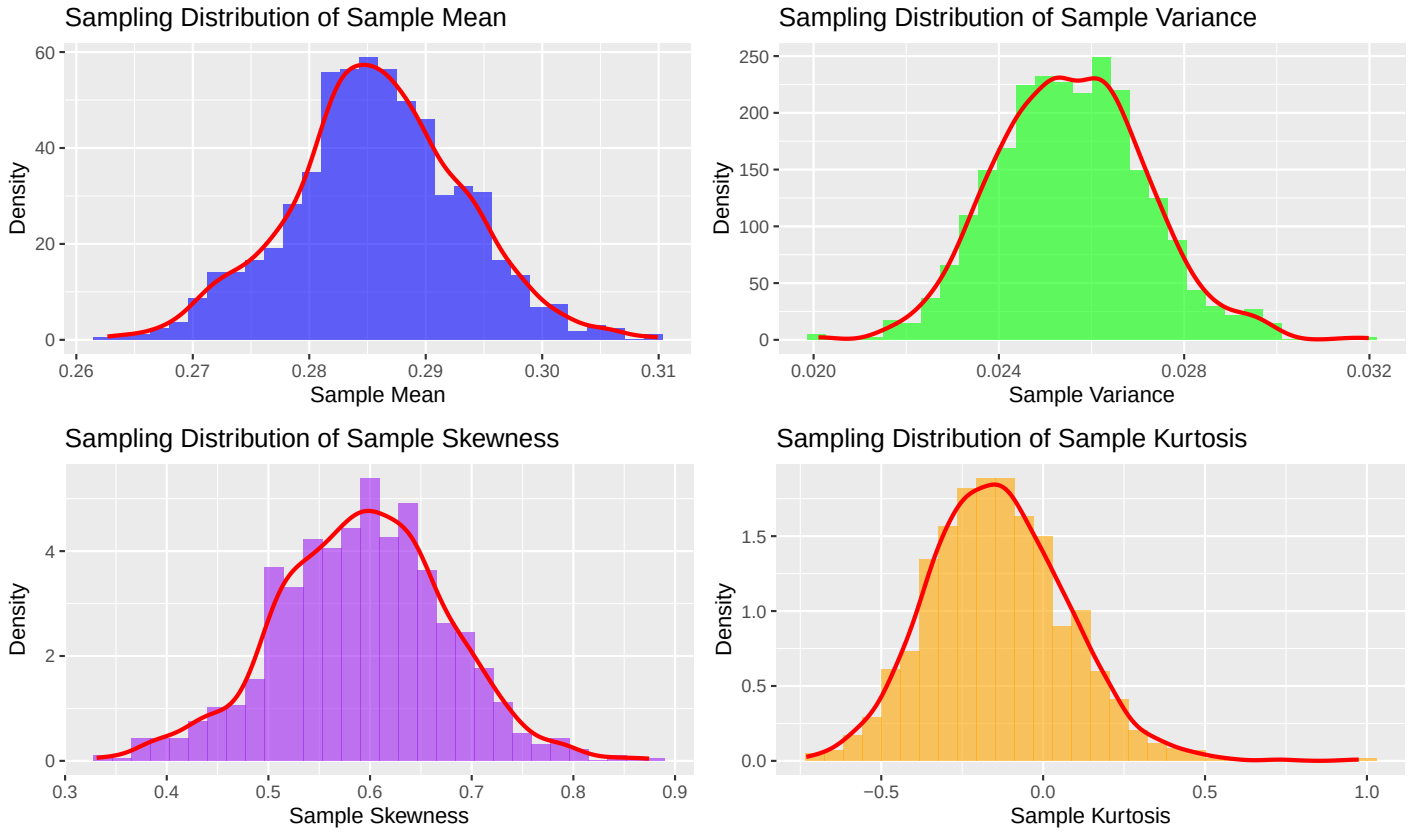

Figure 3: α=2 and β=5 Sampling Distribution Variation

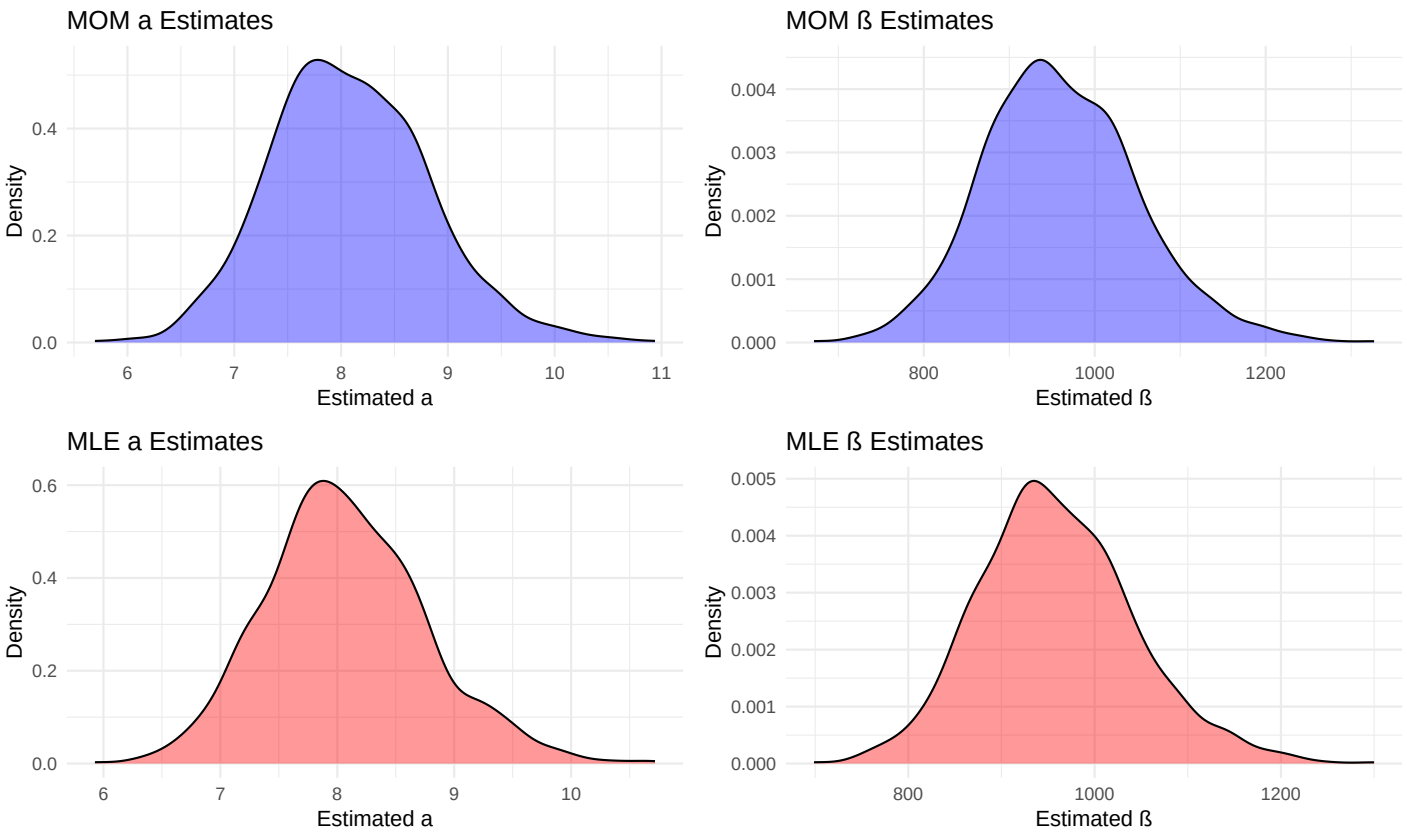

Figure 4: Estimator Distribution