Lab 7 and 8 – MATH 240 – Computational Statistics

Caroline Devine
Colgate University
Math Department
cdevine@colgate.edu

April 1st, 2025

Abstract

This lab evaluates and examines the beta distribution seeing how varying parameters affect its shape and statistical properties, focusing on the mean, variance, skewness, and excess kurtosis. Through analyzing real-world data from World Bank, we see the accuracy of how sample estimates approximate the population parameters with different sample sizes and estimators. Our results show that Maximum Likelihood Estimation provides slightly more precise estimates, but both Method of Moments and Maximum Likelihood Estimation are sufficient estimators.

Keywords: Beta Distribution, Real-World Data, Parameter Estimation, MLE, MOM, Law of Large Numbers

1 Introduction

The beta distribution is a continuous distribution modeling a random variable `X` ranging from 0 to 1. It is extremely useful for modeling proportions, probabilities, or rates. It is also known for being a flexible distribution in regards to its shape, meaning it can exhibit left-skewness, right-skewness, or symmetrical dependent upon its parameters. It is shaped by two parameters, α>0 and β>0. The goal of this lab is to understand the following questions: What is the beta distribution? What does it look like? What is it used for? What are its properties? What information do the simulations and real data analysis provide? Thus, this lab assesses both the theoretical properties of the beta distribution and sees how these properties hold through estimation from real-world data.

2 Density Functions and Parameters

The beta distribution is defined by its probability density function (PDF).

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

The beta distribution's shape is defined by its parameters and thus, the population-level characteristics are described by those parameters.

To explore the beta distribution, we focused on four cases: Beta($\alpha = 2, \beta = 5$), Beta($\alpha = 5, \beta = 5$), Beta($\alpha = 5, \beta = 2$), and Beta($\alpha = 0.5, \beta = 0.5$). We calculated the population-level characteristics(mean, variance, skewness, and excess kurtosis) for all four cases by deriving the formulas and numerically analyzing them for each case (Table 1).

Alpha	Beta	mean	variance	skewness	kurtosis
2.00	5.00	0.29	0.03	0.60	-0.12
5.00	5.00	0.50	0.02	0.00	-0.46
5.00	2.00	0.71	0.03	-0.60	-0.12
0.50	0.50	0.50	0.12	0.00	-1.50

Table 1: Population-Level Summary by Case

To visually see the distribution, all four cases are plotted in Figure 1. The Beta($\alpha = 2, \beta = 5$) distribution is right-skewed, the Beta($\alpha = 5, \beta = 5$) is symmetric, the Beta($\alpha = 5, \beta = 2$) is left-skewed, and the Beta($\alpha = 0.5, \beta = 0.5$) is u-shaped indicating majority of the distribution is close to 0 or 1. We can see from Figure 1 that when $\alpha < \beta$, it is right-skewed; when $\alpha > \beta$, it is left-skewed; when $\alpha = \beta$, it is symmetric.

3 Properties

The beta distribution includes many important properties such as mean, variance, skewness, and excess kurtosis. These values were calculated for all four cases using provided formulas from beta distribution (Table 1). To compare the results, the moments of the beta distribution were calculated directly using numerical integration. Our function, `beta.moment()`, calculated both the centered and uncentered moments, summarizing the same characteristics as in Table 1. This process resulted in values that match the theoretical values from Table 1 and can be seen in Table 3 in Appendix.

Since the goal of summarizing this data is to approximate what the population distribution might be, we further analyze how different sample sizes effect the sample estimates. Initially randomly selecting sample sizes of `n` = 500 from the beta distribution, histograms for each sample with estimated density and true PDF were plotted (Figure 2) accompanied by a numerical summary (Table 4). Comparing numerical summaries and plots from sample to Table 1, confirms that the sample estimates the approximate theoretical values.

With statistics, sample size can heavily influence the accuracy of estimates. When estimating parameters of a theoretical distribution like the beta distribution, the sample size can affect the variability of the numerical statistics. The Law of Large Numbers (LLN) states when the sample size increases, the sample mean converges to the true population mean. This applies to the sample statistics we are looking at. To analyze this, we computed the cumulative numerical summaries for Beta($\alpha = 2, \beta = 5$) data using the `cumstats` package, plotting the cumulative statistics against the true values that describe the actual population distribution seen in Table 1 (Erdely and Castillo, 2017). In Figure 3, when the sample size (`n`) is small, there is more variability in the statistics. As the sample size increases, the statistics converge closer and closer to the true population statistic. This highlights the importance of sample size when calculating samples from a distribution.

Further analyzing the Beta($\alpha = 2, \beta = 5$) distribution, we simulated 1000 samples, computing the mean, variance, skewness, and excess kurtosis to summarize the statistics to see their distribution. Figure 4 shows 4 histograms to visually see the sampling distributions of each statistic. The distribution of the sample means looks to follow a normal distribution with a peak at 0.2875 (the theoretical population mean for Beta($\alpha = 2, \beta = 5$)). The variance follows the same pattern as the mean. The skewness appears normal with variability in spread indicating skewness estimates are more variable with smaller samples. Lastly, the excess kurtosis follows a similar analysis to the skewness. This further emphasizes the importance of sample size.

4 Estimators

Both the parameters of the beta distribution, α and β, are unknown in real-world applications, thus estimation of these parameters are essential in interpreting observed data. The most widely used estimators are the Methods of Moments (MOM) and the Maximum Likelihood Estimation (MLE). MOM equates the sample moments to the theoretical moments of distribution by solving a system of equations to find the parameter estimates. Typically, MOM is easy to compute when the moments of distribution are available in closed form and sample sizes are sufficiently large. MLE, on the other hand, calculates the parameter estimates using the likelihood function for the given distribution, in this case, the Beta distribution, and maximizes it based on the observed data. This is used over the MOM when accuracy is required for observed data of various sample sizes, the likelihood function is not difficult to derive, and for more complex distributions. Computationally, the log-likelihood function is easier to work with using the `optim()` function in `R` to maximize it.

5 Example

Faith (2022) suggested that country death rates worldwide can be modeled with a beta distribution (Fatih, 2022). To see if the beta distribution fits the real-world death rate data, we analyzed the data from the World Bank in 2022 with the rate as deaths per 1,000 citizens. Using MOM and MLE, we estimated the parameters for the death rate data, comparing the MOM and MLE estimates to see which method would provide a better estimate of the beta distribution parameters for this data (Hasselman, 2023). Table 2 shows that both the MOM and MLE provide close estimates. There is a slight increase in variability for the parameter β. Figure 5 shows a visual representation of the fit of both the MOM and the MLE which both indicate strong, reasonable fits to the distribution.

Method	alpha	beta
MOM	8.43	1003.36
MLE	8.27	985.22

Table 2: Parameter Estimations

To analyze the MLE and MOM estimators preformance, we found the MOM and MLE estimates for 1,000 samples with Beta($\alpha = 8, \beta = 950$) and plotted the density estimates for α and β. Coinciding with the plot, the bias, precision, and mean squared error were also calculated for the estimates (Table 5). Figure 6 shows that both the MLE and MOM provide good estimates, but the MLE is slightly more precise.

6 Discussion

Through both theoretical and empirical analysis, we can see how the beta distribution preforms with various data and parameters. The shape of the beta distribution can change, measured by skewness, based on varying the parameters of α and β. We also highlighted how sample size influences the accuracy of estimates of numerical statistics due to LLN. To answer the question of which estimator is more efficient in estimating the parameters, we can see in our real-world example that MLE showed slightly better precision, but both MOM and MLE provided sufficient estimations. Overall, the results of this lab emphasize the flexibility of the beta distribution.

Bibliography: Note that when you add citations to your bib.bib file *and* you cite them in your document, the bibliography section will automatically populate here.

References

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

Fatih, C. (2022). *Determinants of Mortality Rates from COVID-19: A Macro Level Analysis by Extended-Beta Regression Model*. Revista de Salud Pública, 24(2):1.

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

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2024). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-16.

7 Appendix

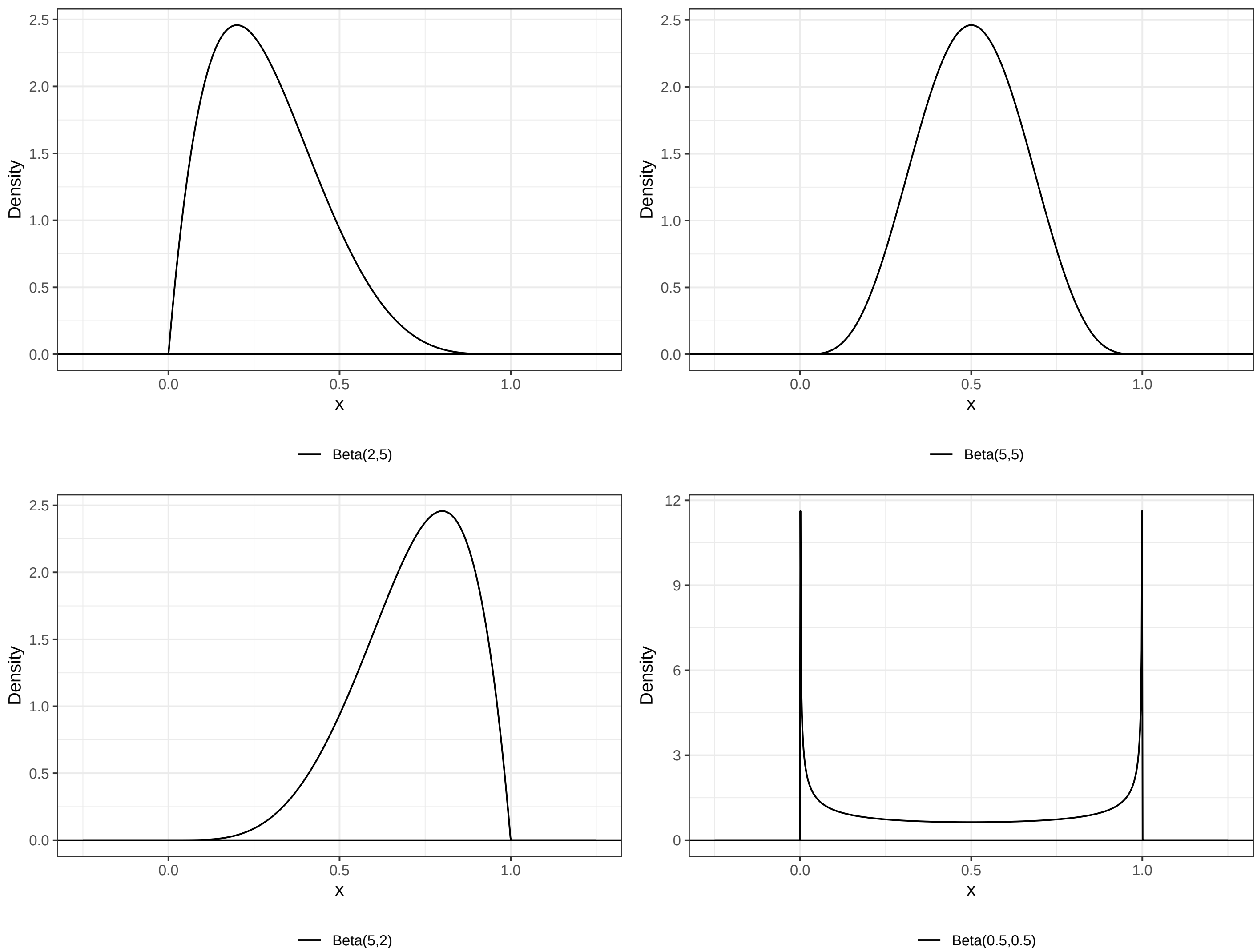

Figure 1: Population Distribution Plot by Case

Alpha	Beta	mean	variance	skewness	excess.kurtosis
2.00	5.00	0.29	0.03	0.60	-0.12
5.00	5.00	0.50	0.02	-0.00	-0.46
5.00	2.00	0.71	0.03	-0.60	-0.12
0.50	0.50	0.50	0.12	-0.00	-1.50

Table 3: Moments Summary Table

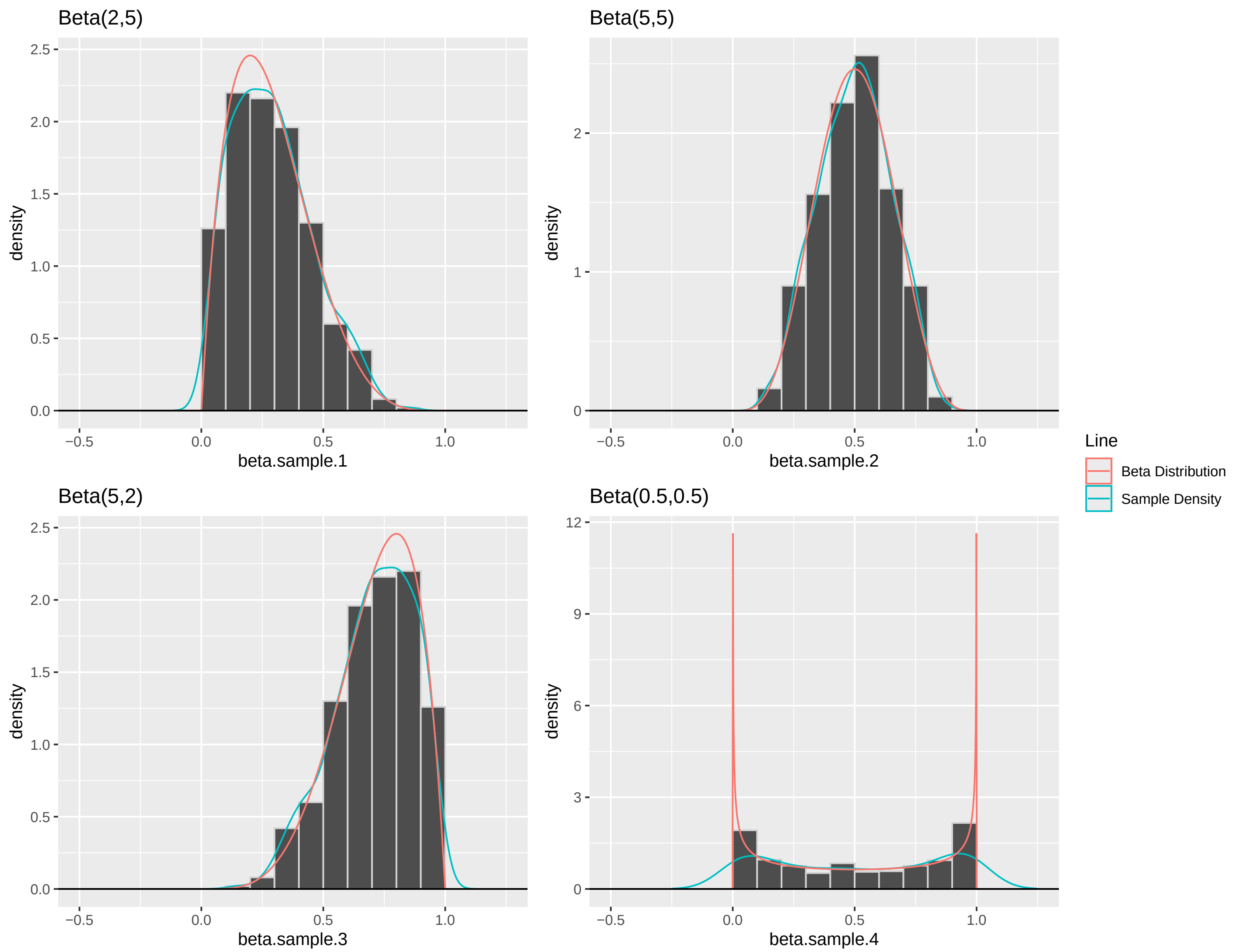

Figure 2: Sample Plot by Four Beta Distributions (Meyer et al., 2024)

Alpha	Beta	mean	variance	skewness	excess.kurtosis
2.00	5.00	0.29	0.03	0.60	-0.12
5.00	5.00	0.50	0.02	-0.00	-0.46
5.00	2.00	0.71	0.03	-0.60	-0.12
0.50	0.50	0.50	0.12	-0.00	-1.50

Table 4: Sample Summary Table

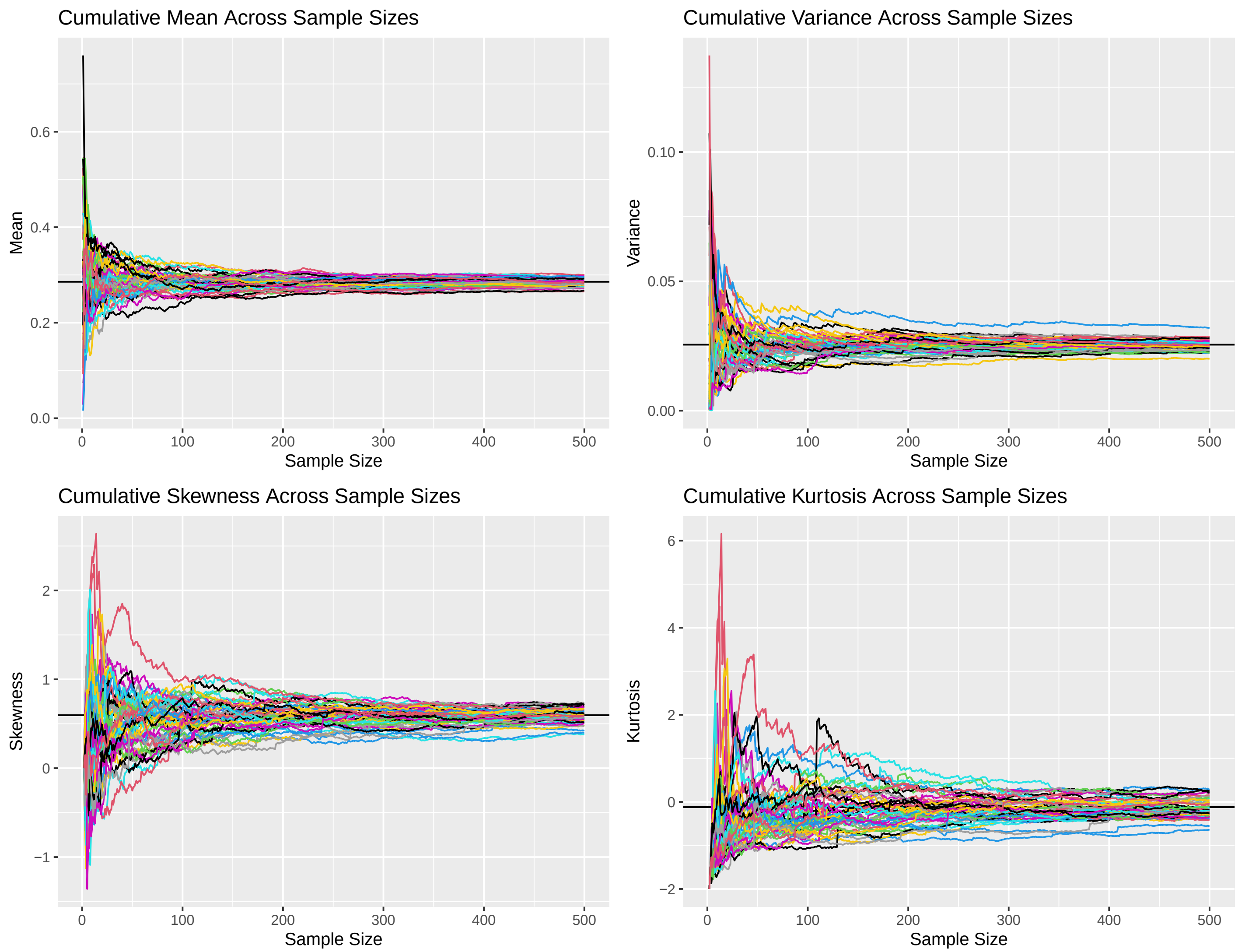

Figure 3: Cumulative Stats For Increasing Sample Sizes

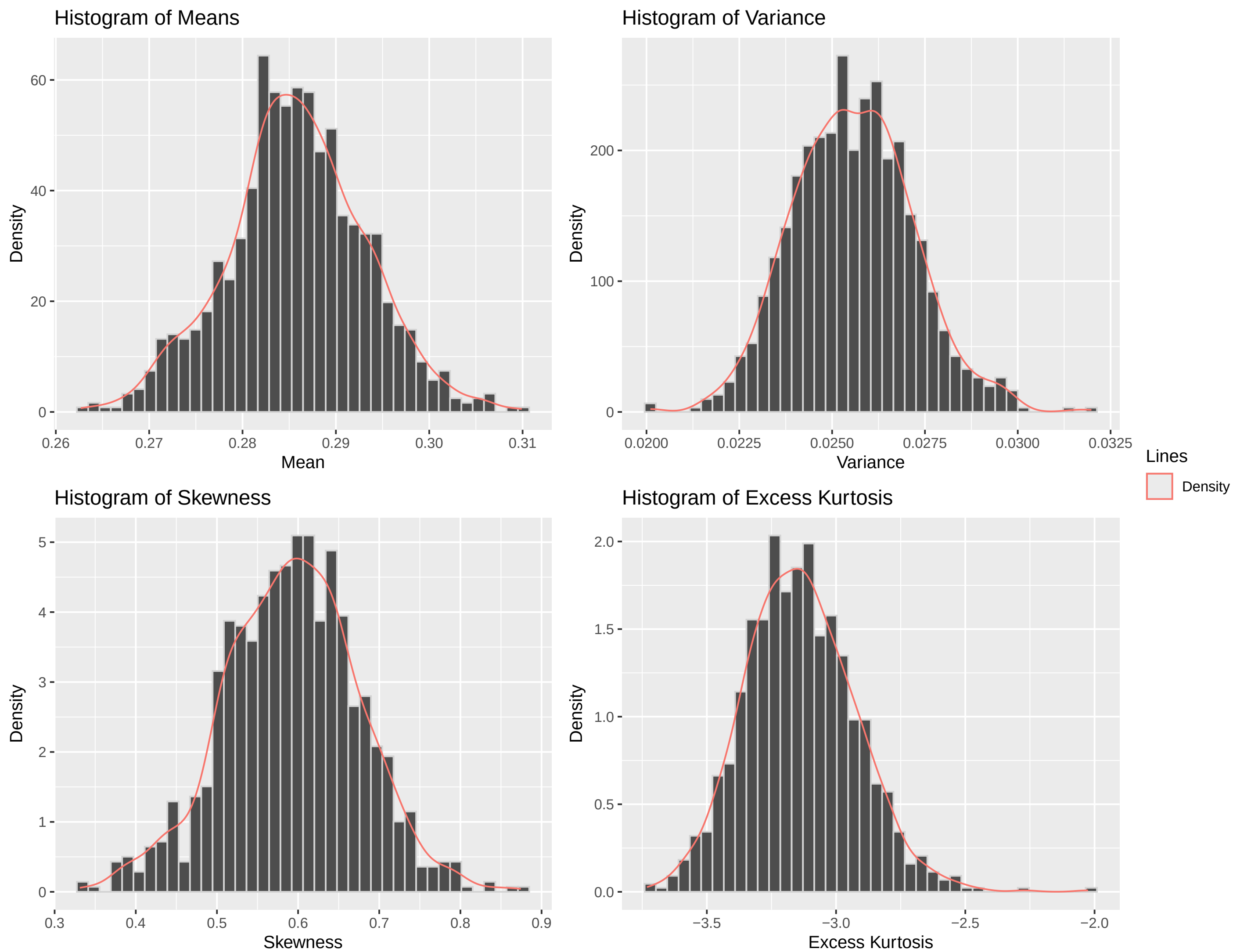

Figure 4: Histogram of Statistics

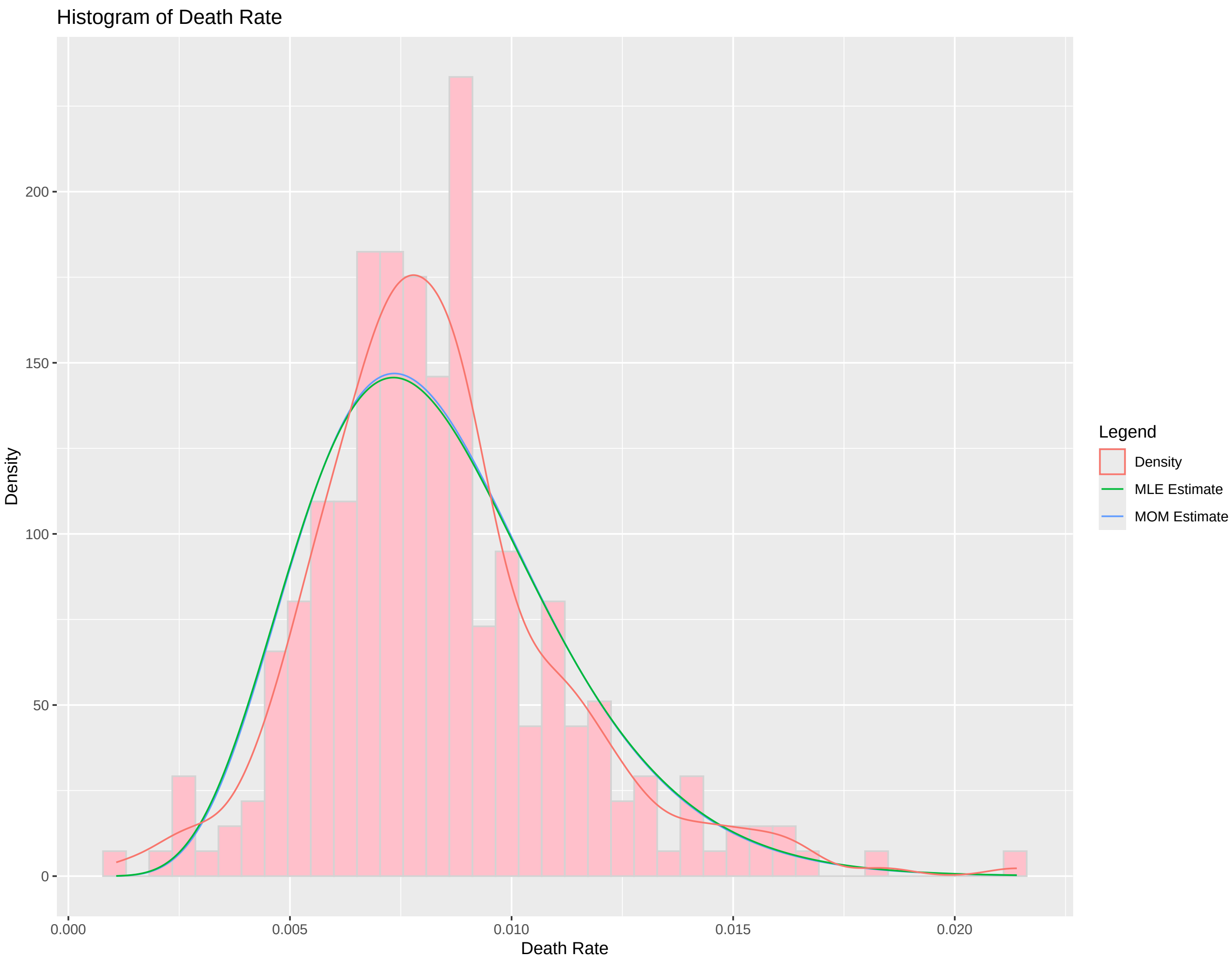

Figure 5: Histogram of Death Rate

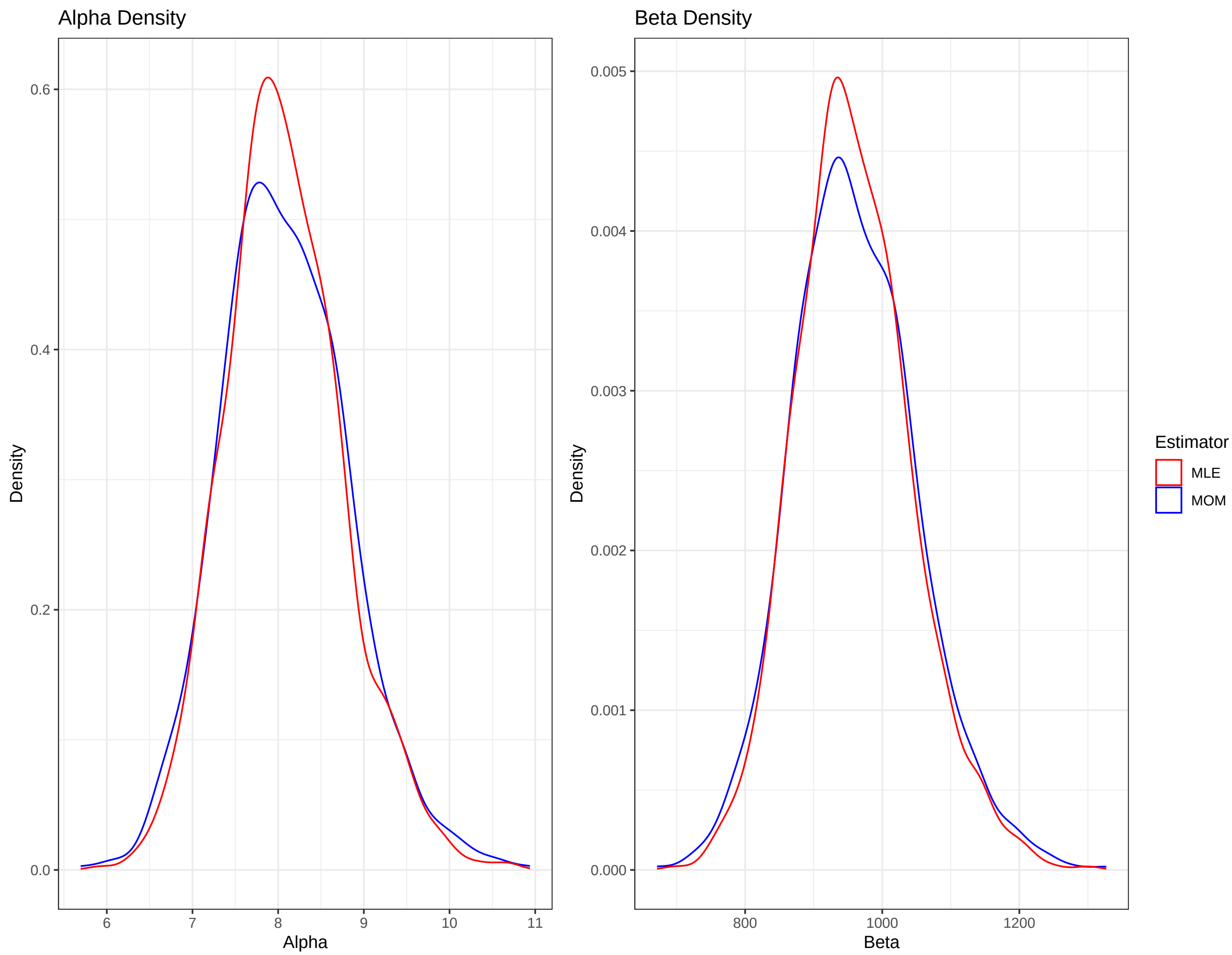

Figure 6: Density for MOM and MLE: Alpha and Beta Estimates

Parameter	Bias_MOM	Bias_MLE	Precision_MOM	Precision_MLE	MSE_MOM	MSE_MLE
Alpha	0.08	0.07	1.83	2.13	0.55	0.47
Beta	10.57	9.18	0.00	0.00	8303.00	7118.73

Table 5: Summary for MLE and MOM