Labs 7-9 – MATH 240 – Computational Statistics

Harrison Wolfe
Colgate University
Math Department
hwolfe@colgate.edu

4/3/2025

Abstract

In this lab we analyzed the beta distribution and its real world application. We attempted to analyze the unique attributes of the beta distribution and then display the results in graphs and tables to show how accurately we can use the beta distribution given a sample or how well we can estimate the population level properties like mean, variance and skewness using estimators.

Keywords: Beta distribution; Estimation; Parameters; Probability Distributions

1 Introduction

The aim of this lab was to analyze several aspects of the beta distribution. We analyzed the parameters, alpha and beta, the properties of the distribution based on a population and sample. We also completed an example of a real world application of the beta distribution discussing death rates in various countries.

2 Density Functions and Parameters

The probability density function or PDF of the beta distribution is based on the parameters alpha and beta. The exact formula for the beta distribution is given below:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt} \cdot I_{(0<x<1)}$$

As you can see in this formula that the graph is almost completely reliant on the parameters of alpha and beta. These parameters will durastically change the distribution, both visually and in summary statistics. Some examples with density functions of different parameters can be seen below in Figure 1:

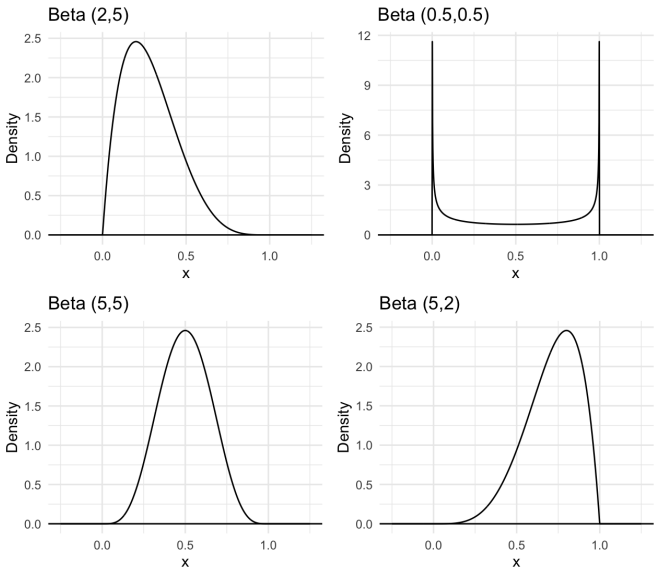

Figure 1: Beta Distributions Given Different Parameters

This graph and all following graphs were made using `ggplot2` and organized using `patchwork` (Wickham, 2016) (Pedersen, 2024). In these graphs we can see how different each PDF of the distribution can be based on the parameters. Even for something that is completely symmetric it can vary a lot. In the (5,5) distribution it is symmetric and unimodal while the (0.5,0.5) distribution is symmetric but bimodal. While some probability functions, like the normal curve's, parameters are more meaningful to the sample describing mean and variance, the beta distributions alpha and beta are not as straight forward. The ways to estimate them will be found below however given a sample it is not as simple to find their parameters with something that describes a direct summary statistics of the sample.

3 Properties

One of the most important way of describing a distribution is through sample statistics. Four of the most important sample statistics are the mean, variance, skewness and kurtosis. When discussing these sample statistics it is important to discuss how sample size affects these various statistics. As we can see in Figures 2 and 3 below each of the sample statistics

approach the population value when the sample size increases. In the first graph we see this for just one sample while in the next we see it for a number of samples showing this reverse megaphone shaped distribution proving that these values approach a certain population value. Another important aspect of sample szie to acknowledge is the large variability that can happen. As we can see with the smaller sample sizes the sample statistics jump all over the place while as it increases it still has some variability but much less. Samples in these graphs were randomly generated samples under the condition that alpha = 2 and beta = 5 using a varying sample size. It is very important to acknowledge these statistics because they give us information like the expected value and how spread apart the data is which is very relevant when thinking about context for real world data following the beta distribution. To compute the skewness and excess kurtosis in these graphs we needed to use the packages `cumstats` and `e1071` (Erdely and Castillo, 2017) (Meyer et al., 2024).

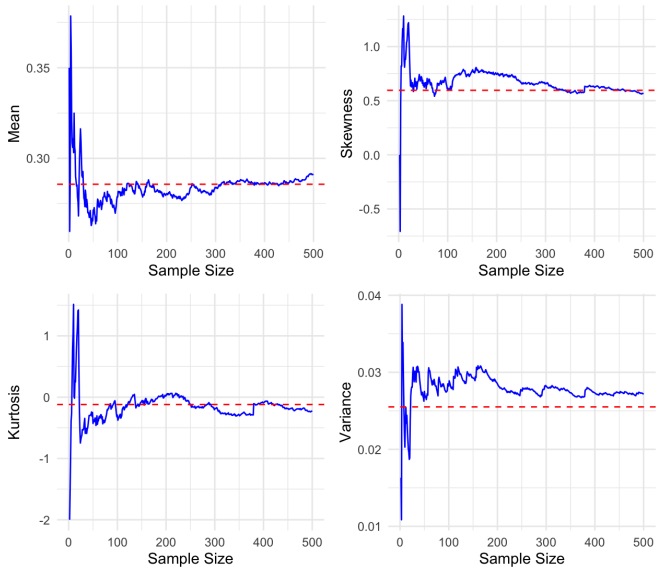

Figure 2: How Sample Size Effects Summary Stats

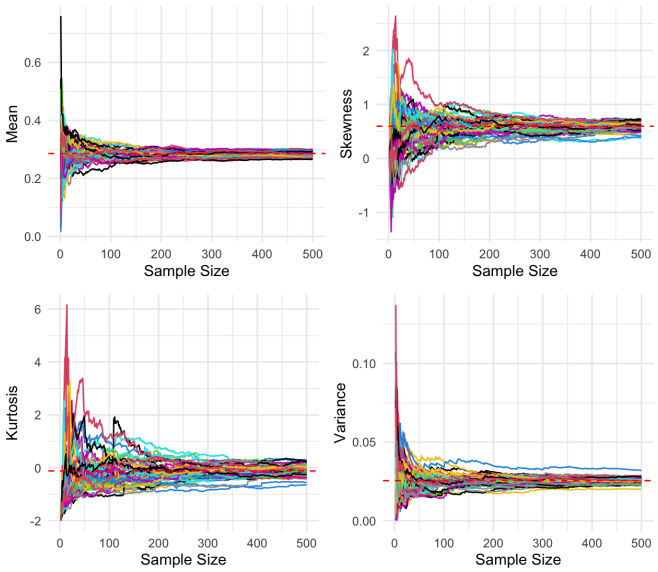

Figure 3: How Sample Size Effects Summary Stats (Simulated)

Figure 4 displays a similar concept about these sampling distributions. They all display that the values in the simulation are distributed with a peak at the population value. In other words, they prove that the population value is what it is. For example in the beta 2,5 distribution the mean in each sample when graphed approached 2/7 or 0.2857.

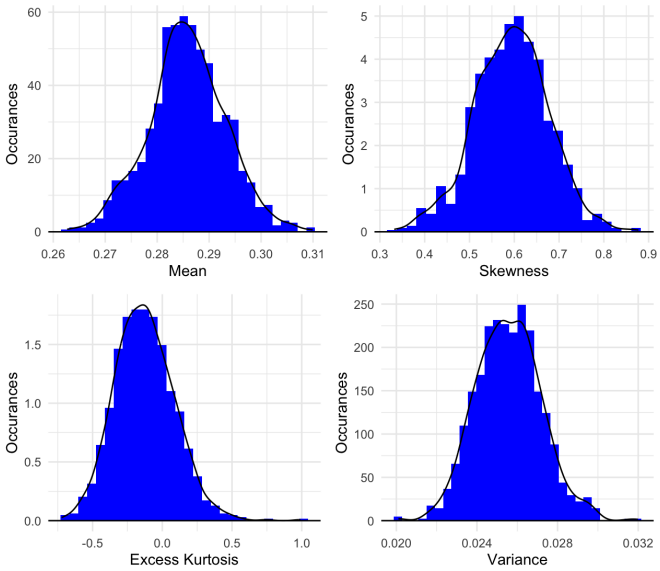

Figure 4: Sampling Distributions of Summary Statistics

Another important aspect of sample statistics is the way of the finding them. This is different for other probability distributions but there are formulas that can be derived for the beta distribution in closed form. This will give exact statistics about the population of the distribution. However, it is important to be able to estimate these statistics for other distributions so we calculated these computationally as well using moments. Moments use the parameters of the graph and take a region under the curve multiplied by a separate

function to estimate a sample statistics like mean. We found each sample statistics using both the formula and in the derived way and this was presented in table 1 in the appendix.

4 Estimators

In order to use the beta distribution we need to be able to estimate the parameters given a sample. There are many ways to do this however two of the most popular are using method of moments and maximum likelihood estimators. The method of moments estimation takes different versions of the sample mean whether it is the mean of the sample itself, the mean of each value squared, the mean of each value cubed, etc. and relates those to the parameters. Since the beta distribution has only two parameters, alpha and beta, we only need two equations to solve. In this case we used the sample mean itself and the mean of each value in the sample squared solving a system of equations. We solved this system using the `nleqslv` package (Hasselman, 2023). For the maximum likelihood estimator we are trying to optimize a product. While this is very difficult using calculus because of the product rule we took the log of the product to turn it into a sum to make the optimization more simple. We then found the values of alpha and beta that maximize the likelihood function giving the maximum likelihood estimators for the beta distribution. In figure 5 below we ran a simulation to compute the maximum likelihood estimators and method of moment estimators for 1000 different random samples given the parameters of alpha = 8 and beta = 950. As we can see in the graph they both converge at those two values. However, picking a better estimator is difficult to see within this graph. If we reference table 2 in the appendix however we can clearly see that the maximum likelihood estimators are more precise for both alpha and beta. This is due to the precision being higher, bias being lower and the mean squared estimate being lower.

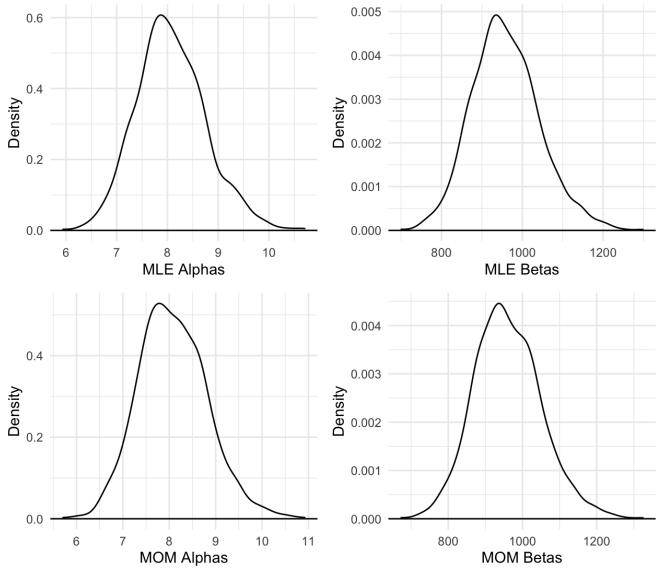

Figure 5: Simulated Estimators for Beta (8,950) Sample

5 Example

The most important aspect of the beta distribution is to be able to apply it to real world situations. An example we observed was the proportion of deaths per 1,000 people in many countries around the world. In this sample we computed both the maximum likelihood estimators and method of moment estimators and fit them to the distribution. In Figure 6 we can see that the MLE and MOM estimators are roughly the same and somewhat accurately describe the distribution. Fitting this sample to a probability distribution will allow us to compute probabilities about the data, simulate future values and is overall very useful. In this particular example we could use the beta distribution to predict the probability of a country having a death rate within a specific range or we could use this data to make assumptions about various regions or groups under a given condition. We were able to organize this data using the `tidyverse` package (Wickham et al., 2019).

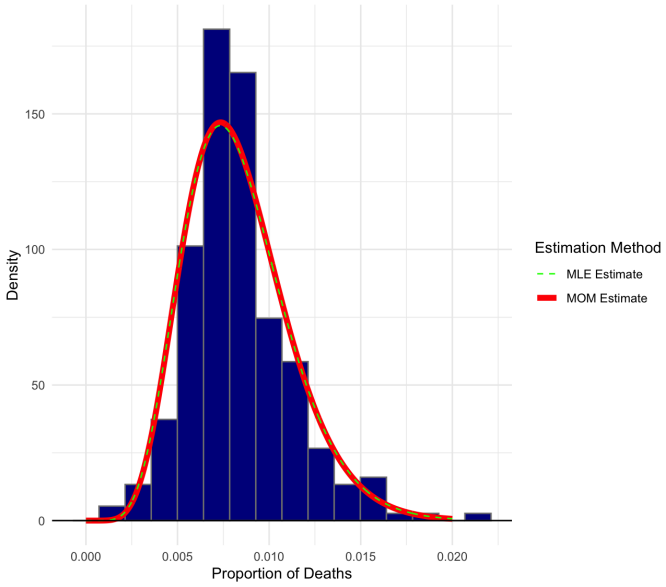

Figure 6: Superimposed Distributions with Estimates on Sample

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

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4.

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.

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.

Pedersen, T. L. (2024). *patchwork: The Composer of Plots*. R package version 1.3.0.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

6 Appendix

These figures could not fit in the margins of the template above so they have been placed here. Each graph was created using `ggplot2` and organized using `patchwork` (Wickham, 2016) (Pedersen, 2024). Each table was made using `xtable` (Dahl et al., 2019).

This graph displays how the population distribution for various beta PDF's describe a random sample generated given the parameters. As we can see the PDF generally follows the density graph for the histogram displaying that large samples described given various parameters replicate the population distribution somewhat well.

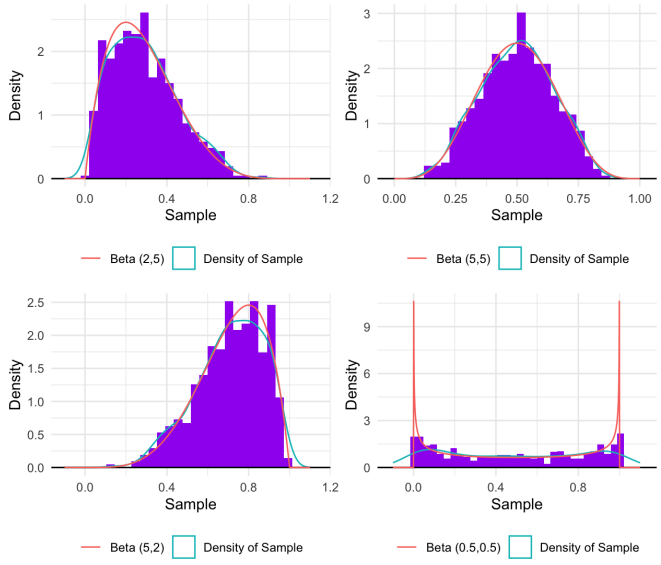

Figure 7: Samples of Beta Distributions

	Alpha	Beta	Mean	Variance	Skewness	Excess Kurtosis	Method
1	2.00	5.00	0.29	0.03	0.60	-0.12	Formula
2	5.00	5.00	0.50	0.02	0.00	-0.46	Formula
3	5.00	2.00	0.71	0.03	-0.60	-0.12	Formula
4	0.50	0.50	0.50	0.12	0.00	-1.50	Formula
5	2.00	5.00	0.29	0.03	0.60	-0.12	Derived
6	5.00	5.00	0.50	0.02	-0.00	-0.46	Derived
7	5.00	2.00	0.71	0.03	-0.60	-0.12	Derived
8	0.50	0.50	0.50	0.12	-0.00	-1.50	Derived

Table 1: Summary Statistics Using Different Methods for the Beta Dist.

	Names	Bias	Precision	MSE
1	Alpha MLE	0.0720	2.1269	0.4754
2	Alpha MOM	0.0822	1.8286	0.5536
3	Beta MLE	9.1142	0.0001	7134.5577
4	Beta MOM	10.3424	0.0001	8290.0516

Table 2: Predictors to determine how good our estimators are