Lab 7-8 – MATH 240 – Computational Statistics

Brendan Mariano
Colgate University
Mathematics
bmariano@colgate.edu

Abstract

This lab provides a template for understanding how a beta distribution works and for how we can find the optimal parameters to model a data set. We made conclusions about the beta distribution through analyzing it with various parameters and we fit it to data about death rates in 2022 using point estimators.

Keywords: beta distribution; density functions; samples; point estimation

1 Introduction

A beta distribution is a tool which is used to model data. The goal of labs 7 and 8 was to learn about how the beta distribution functions and to determine how we can fit it to a data set. In lab 07, we experimented with changing the parameters of a beta distribution, and we also analyzed the distribution of mean, variance, skewness, and kurtosis individually after many samples from a beta distribution. This is discussed in the Density Functions and Parameters section, and the Properties section. Additionally, in order to fit the beta distribution to given data sets, we used point estimators which are discussed in the Estimators section.

2 Density Functions and Parameters

Distribution's are useful because they allow us to model data to the best of our ability, which can allow one to find the probability of certain outcomes. The beta distribution specifically can take several different shapes based on its parameters alpha and beta, but the main shape is a parabolic looking curve that has a maximum. It can also be left skewed, right skewed or normal depending on what the alpha and beta are.

In lab 07, we produced the beta probability density function with four different combinations of parameters for alpha and beta: (2,5), (5,5), (5,2), and (.5,.5). Using the data and ggplot (Wickham, 2016), we produced individual graphs for each set of parameters which also had a normal distribution containing the same mean and variance. Additionally, we included a Gaussian distribution in the plot which contained the same mean and variance as the given Beta distribution. When alpha and beta were equal at (5,5), the beta distribution resembled the normal distribution. When we decreased alpha to two, however, the data became right skewed, and when we decreased beta to two, the distribution became left skewed. Lastly, when we decreased both the alpha and beta to less than one, the distribution was symmetrical but had a greater density on the tails and changed shape overall. Since little data is likely to resemble the last distribution, the beta distribution is most effective when the parameters are greater than one. Ultimately, we can conclude that the distribution is right skewed when alpha is less than beta makes the distribution right skewed, left skewed when alpha is greater than beta, and normal when they are equal.

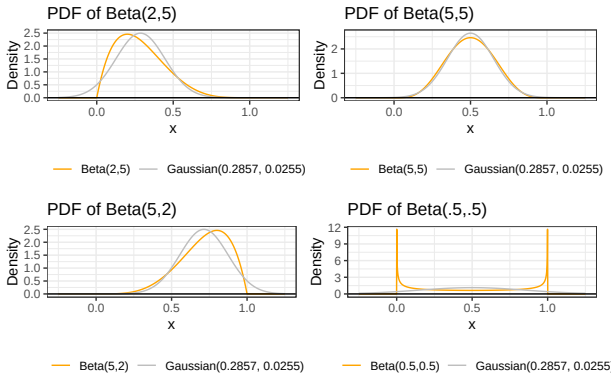

Figure 1

3 Properties

Moments of a distribution are measures for it. There are many of them, but we analyzed four: mean, variance, skewness, and kurtosis. These are calculated using data from a distribution, which in this lab was the beta distribution. Moments can either be un-centered, indicating that they take they don't account for the mean, or centered which means that they account for the mean $(X - \mu_X)$. The equations are as follows.

$$\mu_X = \mathbb{E}(X)$$

$$\sigma_X^2 = \text{var}(X) = \mathbb{E}\left[(X - \mu_X)^2\right]$$

$$\text{skew}(X) = \frac{\mathbb{E}\left[(X - \mu_X)^3\right]}{\left(\mathbb{E}\left[(X - \mu_X)^2\right]\right)^{3/2}}$$

$$\text{kurt}(X) = \frac{\mathbb{E}\left[(X - \mu_X)^4\right]}{\left(\mathbb{E}\left[(X - \mu_X)^2\right]\right)^2} - 3$$

We were able to make a direct comparison between the different sets of parameters by calculating those values and putting

them into a table. One thing that is notable is that the excess kurtosis, variance, and skewness all have the same magnitude whether the parameters are (2,5) or (5,2). The graph for this data can also be seen under the first set of graphs for figure 2.

We then calculated the cumulative moment for each of the four moments for 500 observations, which we obtained randomly from the beta distribution– we repeated this process 48 times. For instance, if we were calculating the cumulative mean, then we would calculate the mean at every observation from zero to 500. We found through our calculations that the cumulative mean of each tract converges to the population (the black y-intercept line) value by 500 observations. This ultimately shows that the beta distribution can be modeled given a hundred or more observations that align with it.

4 Estimators

Although we analyzed the beta distribution in lab 07, we never established a method to find its parameter values for a given data set to apply it. In lab 08, we utilized two estimator methods in order to determine the alpha and beta values for the beta distribution given our data set. The first estimator method is MOM or Measure of Moments. A moment can be calculated based on the population and based on the parameters (alpha and beta). The number of moments that we choose to analyze is equal to the number of unknown variables. The goal of the Measure of Moments is to finds parameter value where the chosen moments are equal in both calculations. The second estimator is the Measure of Likelihood (MLE). MLE is based on the principle of likelihood and uses the probability density function. The probability density function displays the frequency that values occur with respect to other values in the data set or distribution (semantics vary between discrete and continuous data). Given a certain set of parameters, MLE creates a probability density function and uses the function to check the density (amount that values occur relative to others) for each value in the data set. The likelihood is then calculated by multiplying each density together. The parameters that produce the greatest likelihood model the distribution most effectively. One caveat is that we often use the log likelihood because the standard likelihood function becomes extremely close to zero.

4.1 Example

In lab 08, we used the the point estimators with the beta distribution to model data about death rates in 2022,, which we cleaned using Wickham et al. (2019). We plugged in 8 as alpha and 950 as beta for our initial values and then tracked the values that our MLE and MOM functions produced, which allowed us to graph them (figure 4) and make direct comparisons. Just visually you can see that both the MLE and MOM peak at just before 8 for alpha and around 940 for beta; however the MLE estimator method has sharper peaks for both the alpha and beta values, which suggests that MLE is more accurate. This finding is supported by the actual values in table 1. The bias is less than MOM's for both alpha and beta, it is more precise, and MLE has less of an mse. Since the MLE is better with every indicator, it is safe to conclude that it is more effective for representing a population. Although the difference between the MLE and MOM in the population vs estimation graph seems minute, small differences can have a significant impact on large data sets.

$$\text{bias} = \text{mean(distribution)} - \text{actual mean}$$

$$\text{variance} = (\text{standard deviation})^2$$

$$\text{precision} = \frac{1}{\text{variance(distribution)}}$$

$$\text{mean squared error (MSE)} = \text{variance(distribution)} + \text{bias}^2$$

References

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.

5 Appendix

	Element	Bias	Precision	MSE
1	Alpha (MLE)	0.07	2.13	0.48
2	Beta (MLE)	9.11	0.00	7132.70
3	Alpha (MOM)	0.08	1.83	0.55
4	Beta (MLE)	10.29	0.00	8288.46

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

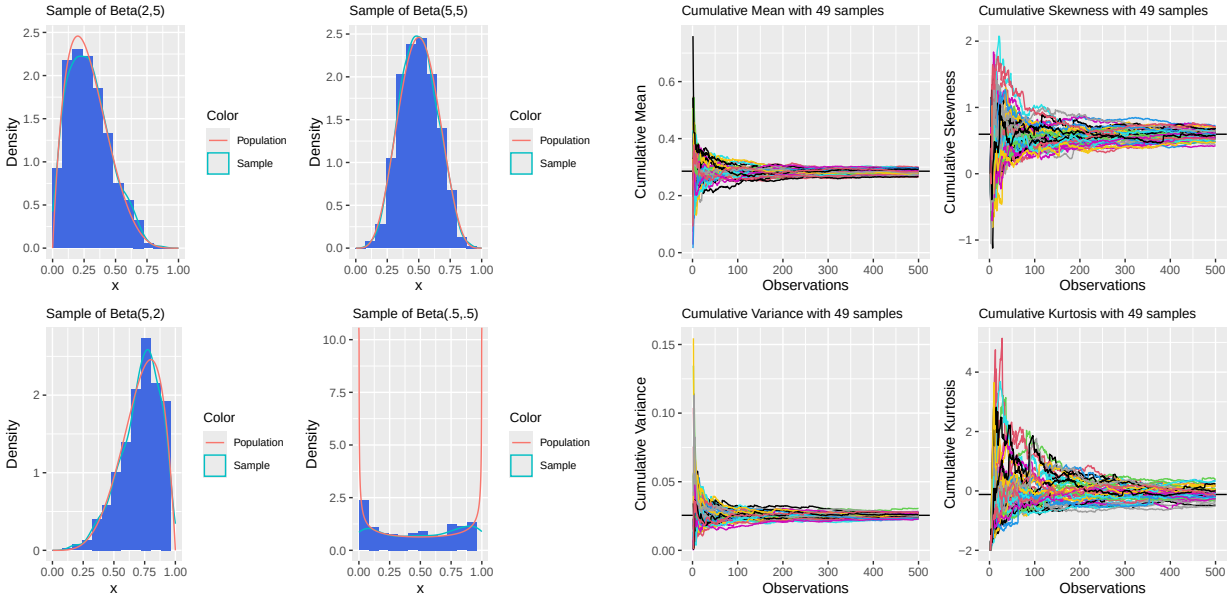

Figure 2

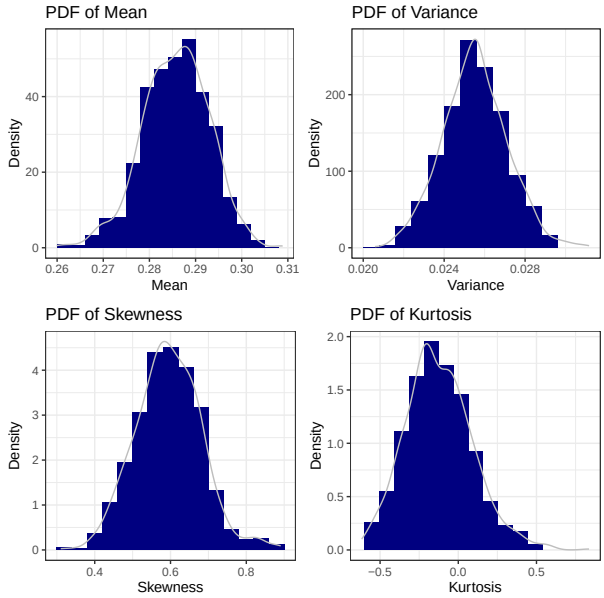

Figure 3

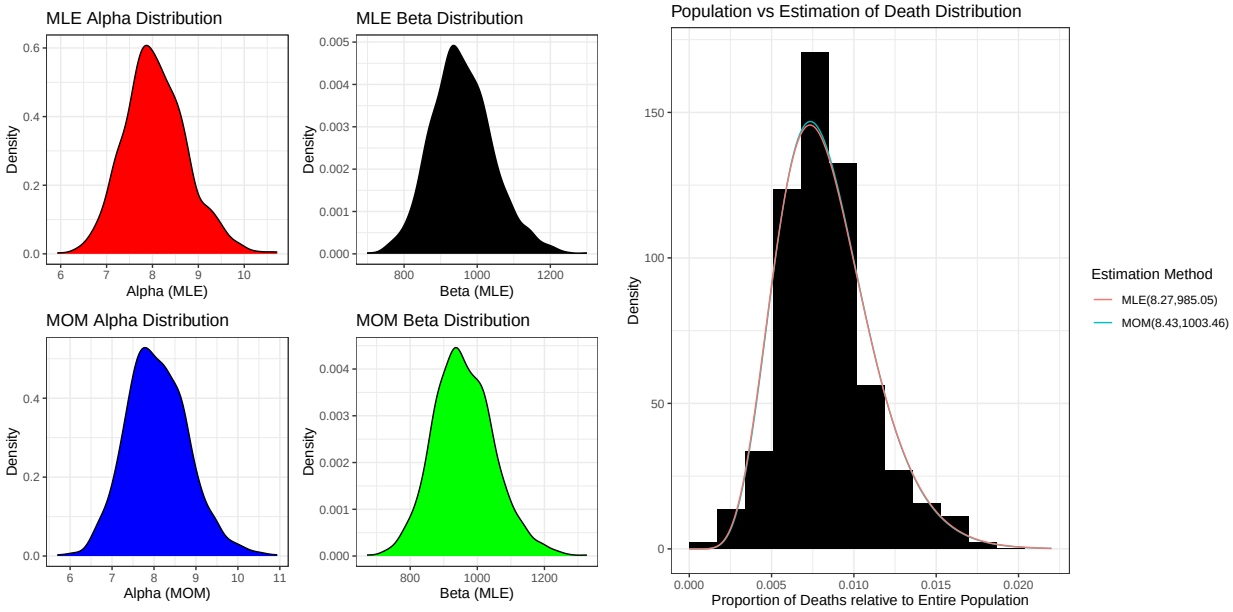

Figure 4