Lab XX – MATH 240 – Computational Statistics

Charles Hooey
Colgate University
Mathematical Economics Major
chooey@colgate.edu

Abstract

This assignment explores the various statistical components beta distribution through the graphical summarization of each component, and analysis of how each population level statistic compares with cumulative statistics sourced from randomly generated data from a known distribution. We then derived the moments of the beta distribution, and used them to compute the population level statistics, and create MOM and MLE point estimator functions. After comparing the graphs and statistics from each parameter obtained from our estimator functions, we concluded that the MLE was the more accurate estimator of the alpha and beta parameters for modeling a distribution with unknown parameters.

Keywords: Point Estimation, Probability Density Function, Parameters

1 Introduction

This laboratory assignment aims to interact with and explore the behavior of the beta distribution. First, we aim to examine how statistical measurements are derived from this distribution, and how their behavior on the population level compares with randomly simulated data from a known distribution with the same parameters. We also seek to better understand the moments of the beta distribution, and how these are used to derive the mean, variance, skewness and kurtosis of a data set. Following this, we seek to determine an efficient way of finding which values of alpha and beta may be used to model a particular set of data, and consequently, determine the most accurate and representative method of point estimation.

2 Density Functions and Parameters

As mentioned before, this laboratory assignment was aimed at deepening our understanding of and strengthening our ability to interact with the beta distribution. Being a continuous distribution, we first interpreted the behavior of the beta distribution through examining how its parameters of alpha and beta dictate the shape of the distribution's probability density function. To do such, we analyzed the beta distribution with four different combinations of the alpha and beta parameters:

- $\alpha = 2, \beta = 5$
- $\alpha = 5, \beta = 5$
- $\alpha = 5, \beta = 2$
- $\alpha = 0.5, \beta = 0.5$

After computing the numerical summary for each of the distributions, we graphed them each with a density plot and superimposed a normal distribution with the same mean and variance to observe the differences between the two as the shape of the beta distribution changes. Comparing all four graphs of the distributions led us to make an assumption that a negative difference between alpha and beta leads to the distribution being more right skewed while a positive difference reflected the opposite. When the two are the same, the data has a skewness of zero. The statistics for each of the parameter combinations are given in the following table:

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

3 Properties of the Beta Distribution

The statistical components of the beta distribution that we analyzed were its mean, variance, skewness and kurtosis. To calculate the statistics of the four pairs of beta distribution parameters we had analyzed numerically and graphically, we first used formulas that were strictly dependent on the parameters alpha and beta. For the second task of this lab, however, we obtained these statistical measurements using the centered and uncentered moments of the beta distribution, which all of the statistics of interest are comprised of. When comparing the values obtained by using the moments to the measurements found with the alpha and beta parameter formulas, we observed that each measurement is almost exactly the same as its respective counterpart.

3.1 Random Data Sampling

The third task of this assignment was to explore how the population statistics of our data would compare with statistics drawn from a generated and known distribution. To accomplish such, we used R to generate a sample beta distribution of size n = 500 for each of the four parameter combinations we were originally working with. The statistics calculated for each of the distributions were close but not the exact same as

our original population level. Furthermore, the plots shared the same visual features as the ones that were created within task one, as the skewness and mean of each were close to one another for each respective combination of parameters.

3.2 Behavior of Cumulative Statistics

Task four of this lab involved examining the behavior of cumulative statistics across many different generated beta distributions of the same parameters (alpha = 2, beta = 5). We used loops and the `cumstats` package for `R` to accomplish this, then plotted the original beta distribution and superimposed lines of different colors for each randomly generated statistic (Erdely and Castillo, 2017). After observing the graphical behavior of each of the statistics, the randomly generated beta distribution statistics suggest that the cumulative statistics for this distribution will all converge to the same value as one another. Task five additionally had us use loops to generate 1000 random samples of size n = 500 for the beta distribution and to graphically summarize each statistic by plotting their distributions. Each of the statistics' distributions were centered almost exactly at their respective population statistic values, which suggests that the cumulative statistics for randomly generated samples should converge to the population level statistics if the sample size is adequate.

4 Point Estimation

Task six of this assignment had us explore how country death rates could be modeled with the beta distribution. Because the support of the beta distribution was 0 to 1, it was first necessary to use the `Tidyverse` package for `R` to wrangle the data and convert each value into a rate over 1000 (Wickham et al., 2019). following such, we created two functions that would estimate the values of alpha and beta through two different types of point estimation: Method of Moments and Maximum Likeliness. After creating a function for and plotting the distributions resulting from the MOM and MLE point estimators, the results both were concurrent with the histogram of the data, and were almost identical with one another, as seen in figure 1. This suggests that both methods of point estimation provided values for alpha and beta that were close to the true values for the distribution of the data.

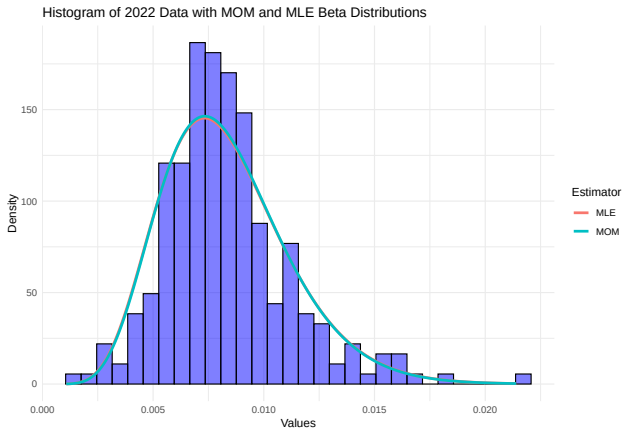

Figure 1: Histogram of Death Rates with MOM and MLE Distributions Superimposed

4.1 MOM and MLE Estimators

After working with the death rate data provided, we pivoted to once again use loops to simulate data, this time in order to compare the distribution of parameters obtained from each method of point estimation. For these calculations, we used values of alpha = 8, beta = 950, and n = 266. In order to make a conclusion on the best estimator for the alpha and beta parameters, we graphed the distribution for each parameter, and calculated their respective bias, precision, and mean square error. Comparing the results from the MOM and MLE estimators provided evidence towards MLE being the more accurate point estimator, as it held a lower bias and higher precision than the MOM, and visually, as observed within figure 2, appeared to follow the histogram of the data more closely than the plot of the MOM distribution did.

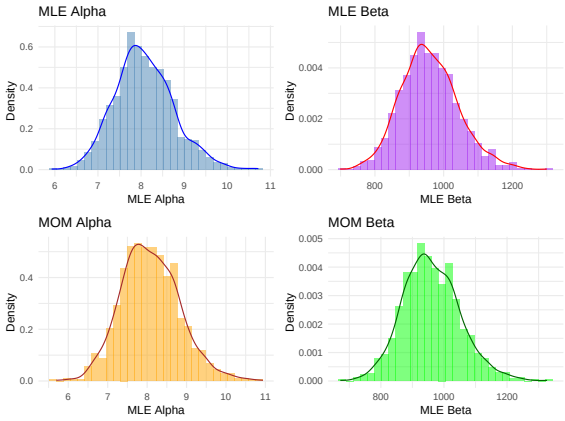

Figure 2: Estimated Probability Distributions for MOM and MLE Parameters

Bibliography:

References

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

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.