Lab 07 – MATH 240 – Computational Statistics

Ben Horner
Colgate University
Math Department
bhorner@colgate.edu

April 1, 2025

Abstract

The beta distribution is useful in modeling probabilities, rates, and other proportion as it and it's key characteristics (mean, variance, skewness, excess kurtosis) are easily defined and calculated. However, when taking a sample of random data, we must estimate those key characteristics. Here, we seek to examine those estimations and compare them to the population-level statistics, find that an increase in sample size will result in the estimated value approaching the population-level. On top of that, the sampling distributions follow a normal distribution around the population-level value.

Keywords: Beta Distribution; Probability Density; Moments; Sample Size; Variation

1 Introduction

The beta distribution is a continuous distribution that is used to model a random variable X that ranges from 0 to 1. This makes it useful for modeling proportions, probabilities, or rates. The beta distribution is also known for being 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: $\alpha > 0$ and $\beta > 0$. We can use these parameters alone to not only define the distribution, but calculate the mean, variance, skewness, and excess kurtosis. Additionally, we can use the centered and uncentered moments of the beta distribution is another way to calculate these population-level characteristics.

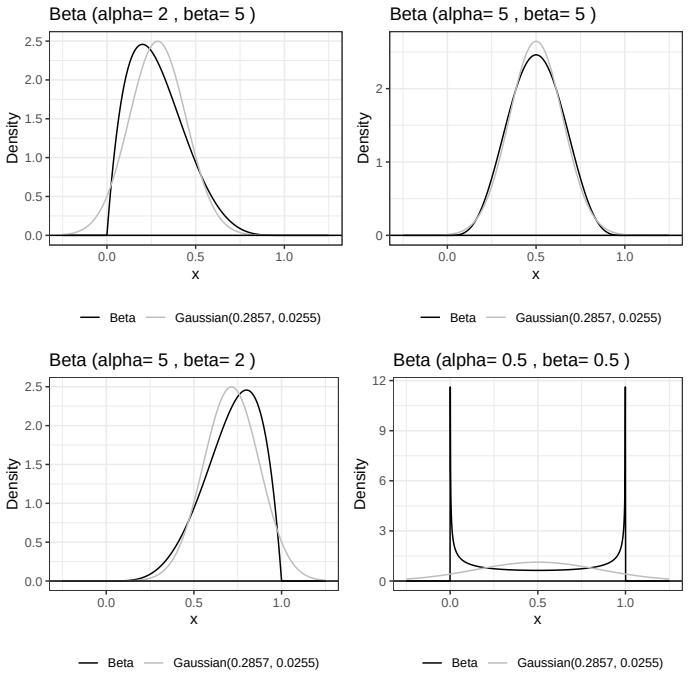
However, often times we may not have access to the beta, alpha, or moment of the distribution, and thus need to approximate it based off of the data. As such, we will examine whether data summaries help, is the sample size important, and also model the variation of the data summaries when compared to the population-level values.

2 Methods

The mean, variance, skewness, and excess kurtosis of a distribution are key summary characteristics that tell us more about the data we are looking at. With a beta distribution, we can calculate the population-level values using either the α and β or the moments. Throughout this process, we load and use the following libraries: `ggplot2` (Wickham, 2016) and `patchwork` (Pedersen, 2024) for plotting and visualizing data, `tidyverse` (Wickham et al., 2019) to manipulate and summarize data, and `e1071` (Meyer et al., 2024) and `cumstats` (Erdely and Castillo, 2017) for additional statistical functions and analysis.

2.1 Describing and Summarizing The Population

As the beta distribution's probability density function is defined in terms of the parameters α and β, the population-level characteristics are similarly defined by them. Using a function we created to summarize the mean, variance, skewness, and excess kurtosis (hereafter referred to as key characteristics) and a function to plot a beta distribution, compared to a Gaussian, we considered four cases: Beta($\alpha = 2, \beta = 5$), Beta($\alpha = 5, \beta = 5$), Beta($\alpha = 5, \beta = 2$), and Beta($\alpha = 0.50, \beta = 0.50$). To confirm our function works, we also computed the population-level key characteristics via the moments of the beta distribution, which match those using alpha and beta.

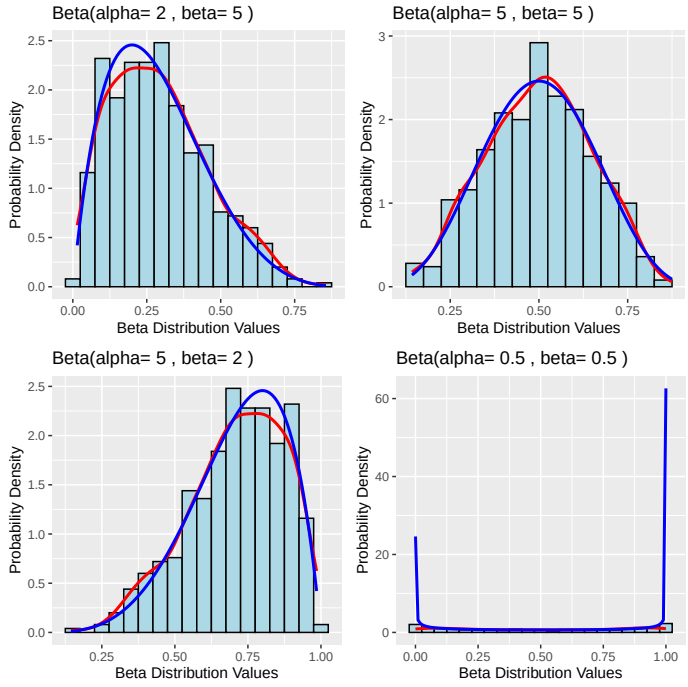


	beta_distribution	mean	varience	skewness	excess_kurtosis
1	2,5	0.29	0.03	0.60	-0.12
2	5,5	0.50	0.02	0.00	-0.46
3	5,2	0.71	0.03	-0.60	-0.12
4	0.5,0.5	0.50	0.12	0.00	-1.50

Table 1: Summary of Beta Distribution Parameters

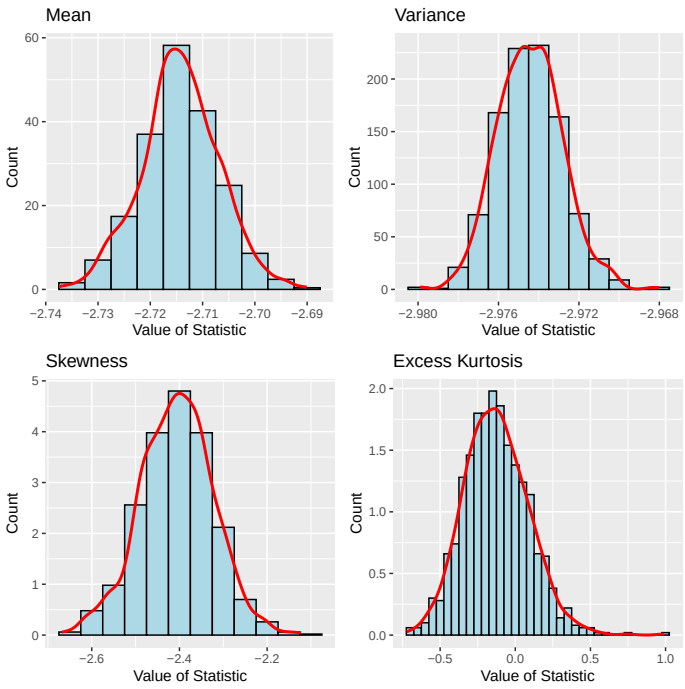
2.2 Summarizing Random Data

When summarizing data, our goal is to approximate what the population distribution might be. We test this by generating a sample of $n = 500$ from each of the considered beta distributions using `set.seed(7272)` to ensure consistency across samples and attempts. We summarize the data using `summarize()` from `dplyr` (Wickham et al., 2019) and plot a histogram of the generated sample and its estimated density (red) compared to the population level beta function for the same alpha and beta (blue).



2.3 Modeling Variation in Data Samples

When taking a random sample of data, each sample varies from the last. As our goal is to model sampled data with a population-level distribution, we examine how our estimations of the key characteristics vary from sample to sample. Beginning with our set seed of 7272 to generate the random sample, we iterate $(1 : 1000)$ through seeds 7272+i, where i is the iteration. Our result is a sample of $n = 1000$ means, variances, skewnesses, and excess kurtosises, which we plot to examine their distribution.



3 Applications: Country Death Rates

Applying the beta distribution in a practical sense, Faith (2022) suggests that country death rates worldwide can also be modeled with a beta distribution. Focusing on data from the World Bank, we use the 2022 data to compute the method of moments estimates and maximum likelihood estimates for the α and β.

	parameter	MoM	MLE
1	alpha	8.08	8.27
2	beta	931.93	985.05

Table 2: alpha and beta of the mom and mle

We now must decide which estimator we should use, so to check the bias, precision, and mse of these estimators, we simulate new data $(n = 266)$ with $\alpha = 8$ and $\beta = 950$. The result is a sample of $n = 1000$ method of moments estimates and maximum likelihood estimates for α and β.

	MLE alpha	MLE beta	MoM alpha	MoM beta
Bias	0.07199	9.11	0.08169	10.2867
Precision	2.12738	0.00014185	1.82856	0.00012221
MSE	0.47524	7132.702	0.55355	8288.461

Table 3: Comparison of MLE and MoM for alpha and beta

4 Discussion

When analyzing data, the key characteristics can provide essential summaries to guide one's research. However, often we only have a sample of the population instead of access to the

true parameters of a distribution and as such, we estimate the key parameters. These summaries do help, and especially as the sample size increases, it approaches the true values for the population. Additionally, the sampling distribution of the beta function's key characteristics appear to follow normal (gaussian) distribution itself, with the standard deviation and spread of the sampling distribution increasing from the mean to varaiance, to skewness, and finally excess kurtosis. Finally, we observe that the MLE is in general, less biased with a lower mse when compared to the MoM. However, the MoM appears to be more precise.

References

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

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.