Lab 8 – MATH 240 – Computational Statistics

Andrew Li
Colgate University
Mathematics Department
ali@colgate.edu

Abstract

In this lab, we explored the beta distribution has been explored thoroughly by working with its various properties, probability distributions, and parameters. By changing the parameters, we can evaluate the effects on its statistical values such as mean, variance, skewness, and excess kurtosis. To analyze real world data on global deaths from the World Bank, we made 2 point estimators (Method of Moments and Maximum Likelihood Estimations), which both work well but the *MLE* works slightly better.

Keywords: point estimations; parameters; probability distributions;

1 Introduction

The beta distribution is a continuous distribution that can be used to model the variability of a random variable X that ranges from 0 to 1. It is useful for modeling proportions, probabilities, or rates as its statistical characteristics are versatile enough to assume many different shapes based on its input parameters (Given that $\alpha > 0$, $\beta > 0$).

By exploring its properties, the effects of various inputs can be seen to answer our questions about what the beta distribution is, what it can be used for, what are some of its properties, and what useful inferences can be drawn from simulation and real data analysis.

The `R` packages that were used are tidyverse(Wickham et al., 2019) for data cleaning and plotting, patchwork(Pedersen, 2024) for combining graphs, e1071(Meyer et al., 2024) for calculating properties, xtable(Dahl et al., 2019) for table creation, nleqslv(Hasselman, 2023) for point estimation calculations, and cumstats(Erdely and Castillo, 2017) for measuring cumulative statistics.

2 Density Functions and Parameters

The beta distribution has a probability density function defined as:

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

Knowing that x stays within 0 to 1, we looked at the different cases of $Beta(\alpha, \beta)$ where $Beta(2,5), Beta(5,5), Beta(5,2),$ and $Beta(0.5,0.5)$ and explored their properties.

The beta distributions for each plot is graphed together in Figure 1.

3 Properties

We calculated the population moments using numerical integration shown in Table 1.

Since the distribution's shape is affected by its parameters, the population characteristics are also controlled by them. To prove that its possible to approximate what the population distribution might be, we can connect our numerical summaries and graphs to the actual distribution by generating random data and comparing the calculated results against those from the known distribution, shown in Figure 2. The properties of mean, variance, skewness, and excess kurtosis are then compared against those from the population characteristics in Table 1.

Then we explored how the law of large numbers is proven to be true as the increasing sample size decreases the variability in the different properties of the data across different samples. This is shown in Figure 3 and remains true when we ran random samples and the graphical representations of the properties end up converging towards the population values as sample size increased.

	variable	mean	variance	skewness	kurtosis
1	Beta(0.5,0.5)	0.50	0.12	0.00	-1.50
2	Beta(2,5)	0.29	0.03	0.60	-0.12
3	Beta(5,2)	0.71	0.03	-0.60	-0.12
4	Beta(5,5)	0.50	0.02	0.00	-0.46
5	Sample Beta(0.5,0.5)	0.52	0.12	-0.11	1.55
6	Sample Beta(2,5)	0.29	0.03	0.57	2.78
7	Sample Beta(5,2)	0.71	0.03	-0.74	3.22
8	Sample Beta(5,5)	0.50	0.02	0.06	2.54

Table 1: population moments

4 Estimators

We created a Method of Moments (MOM) point estimator and Maximum Likelihood Estimator (MLE) to calculate the two unknown parameters of α and β. To use the MOM, we had to find the first two moments of the beta distribution . The first moment is calculated as $\frac{\alpha}{\alpha+\beta}$ while the second is denoted by $\frac{\alpha*(\alpha+1)}{(\alpha+\beta+1)*(\alpha+\beta)}$. Normally, you can create a symbol of equations and, through substitution, find out what each moment would be using only the sample value for the MOM, and the MLE would have required taking the likelihood of every x and then optimizing the values to find the maximum. We reduced the computations required with `R`.

5 Example

After running a thousand samples of size 266 with $\alpha = 8$, $\beta = 950$ to model the world death data, we were able to compare the accuracy (bias) and variability (precision) of the two point estimators. Looking at the Figure 4, it can be seen that the MLE has less variability because its values doesn't spread as much and has a taller peak. This is further verified when looking at the numerical values in table 2, so the MLE is the better point estimator.

bias	precision	mse	names	actual	estimated
0.08	1.83	0.55	moms alpha	8.00	8.08
10.29	0.00	8288.46	moms beta	950.00	960.29
0.07	2.13	0.48	mles alpha	8.00	8.07
9.11	0.00	7132.70	mles beta	950.00	959.11

Table 2: Table to compare estimator values

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., 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

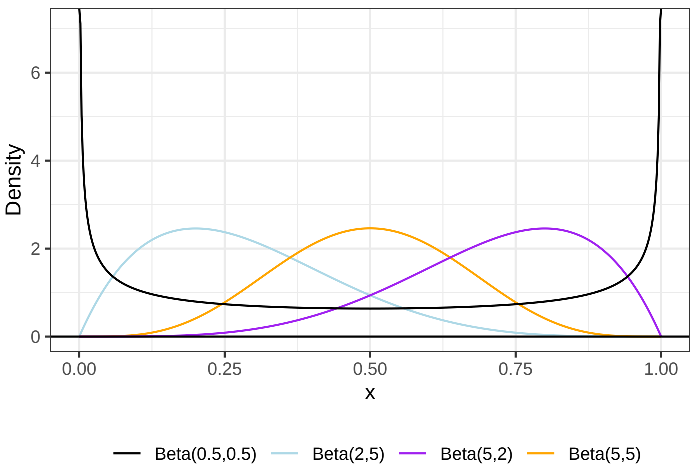

Figure 1: Distributions of different beta plots

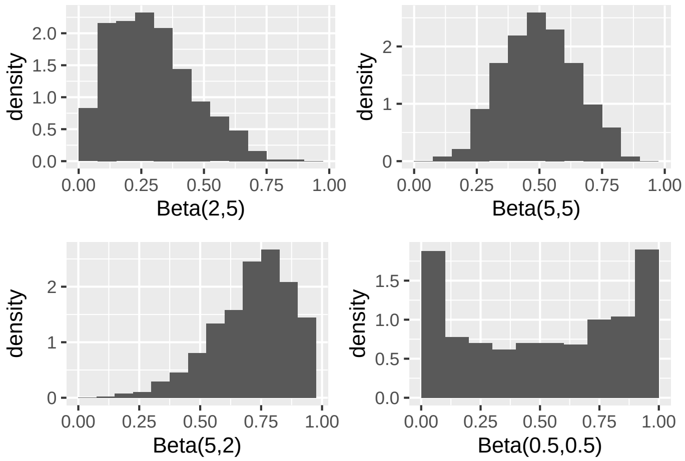

Figure 2: Histograms of densities of beta samples

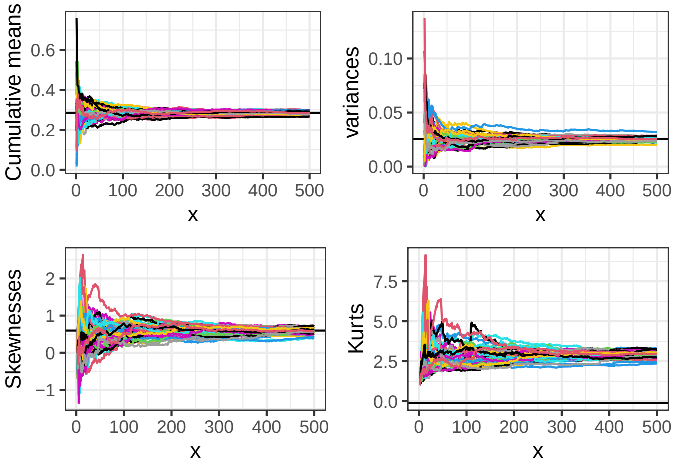

Figure 3: Graphical comparison of random samples to population data

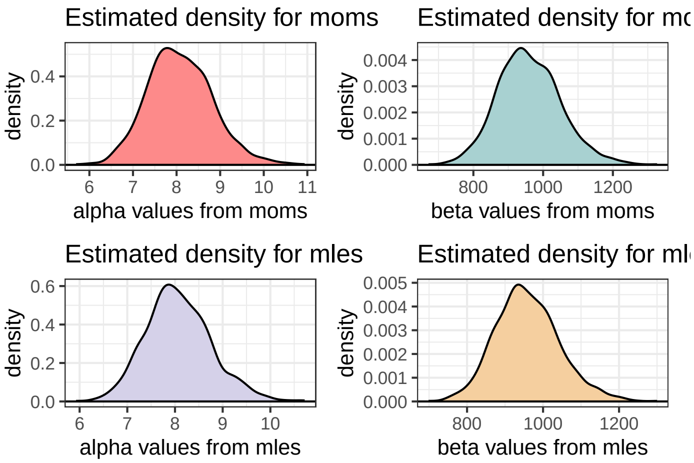

Figure 4: Densities of alpha and beta values from point estimators