Performance of Different Parametric Estimation Techniques with Missing Data

Introduction

In modern statistics, as datasets with missing data become increasingly common, it is important to develop research on analytical problems that may arise when there is missing data. Thus, we want to compare the performance of different parametric estimation techniques under data with different distributions and various percentages of missing data.

This topic is especially relevant in today's world as the majority of real-world datasets are incomplete in some form or another. Since we have spent a lot of time on different techniques for estimating the parameters of the underlying distribution of different datasets in this course, it would be interesting and appropriate to examine how these techniques hold up to real-world data. Therefore, in this paper, we will focus on the performance of the method of moments estimation and the maximum likelihood estimation techniques on the Beta distribution.

Methodology

First, we created three populations of 10,000 observations, each from a different distribution generated by the two-parameter Beta distribution such that each one exhibits a different level of skewness. The two-parameter Beta distribution is characterized by its pdf $x^{\alpha-1} \cdot (1-x)^{\beta-1} / B(\alpha,\beta)$ on support $[0,1]$, where α and β are the shape parameters, and $B(\alpha,\beta) = \Gamma(\beta) + \Gamma(\alpha) / \Gamma(\beta+\alpha)$ is the Beta function. The three distributions from which we created our populations are described below:

1. Normal: produced by the parameters shape $\alpha = 5$ and $\beta = 5$
2. Slightly Skewed unimodal: produced by the parameters $\alpha = 3$ and $\beta = 6$
3. Exponential-like: produced by the parameters $\alpha = 1$ and $\beta = 5$

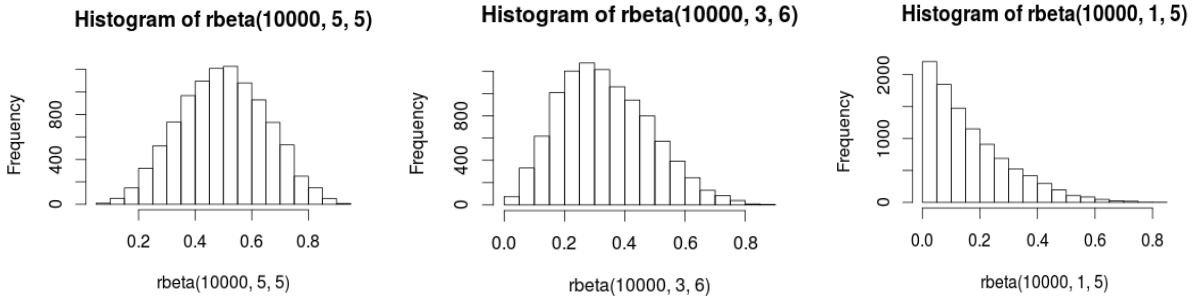

Figure 1. Distributions generated using R Studio.

We used the Beta distribution because it is relatively simple and flexible, being able to take on characteristics of other types of distributions depending on its parameters. This allows us to simulate circumstances where we might have outliers in one or both directions.

Then, to simulate the loss of information when gathering outside data, we randomly generated nine samples for each distribution differing by how data points are removed from the population. Two factors were considered in our study: pattern of loss, and the amount of loss. The three possible amounts of missing data that we considered were 25%, 50%, and 75%, and the possible patterns of loss are:

1. Loss of the first quartile of data
2. Loss of the third quartile of data
3. Random loss of data

In particular, loss from one particular extreme would simulate scenarios where detection or response rates are directly related to the response, such as an instrument outside of its detection range or a nonresponse bias in a survey.

Finally, we used two techniques to estimate the parameters of the sampled datasets:

1. Method of moments estimation,
2. Numerical estimation of the maximum likelihood method.

Iterating this random generation of samples 100 times, we computed the mean and variance of each estimated parameter. From this, we calculated the percentage difference between the original and predicted estimates. Additionally, this also allows us to use a one-sample t-test to calculate how different the new estimated parameters are from the ones for the original distribution, as a measure of the significance of this difference.

Results

The following scatter plots summarize our findings for each distribution.

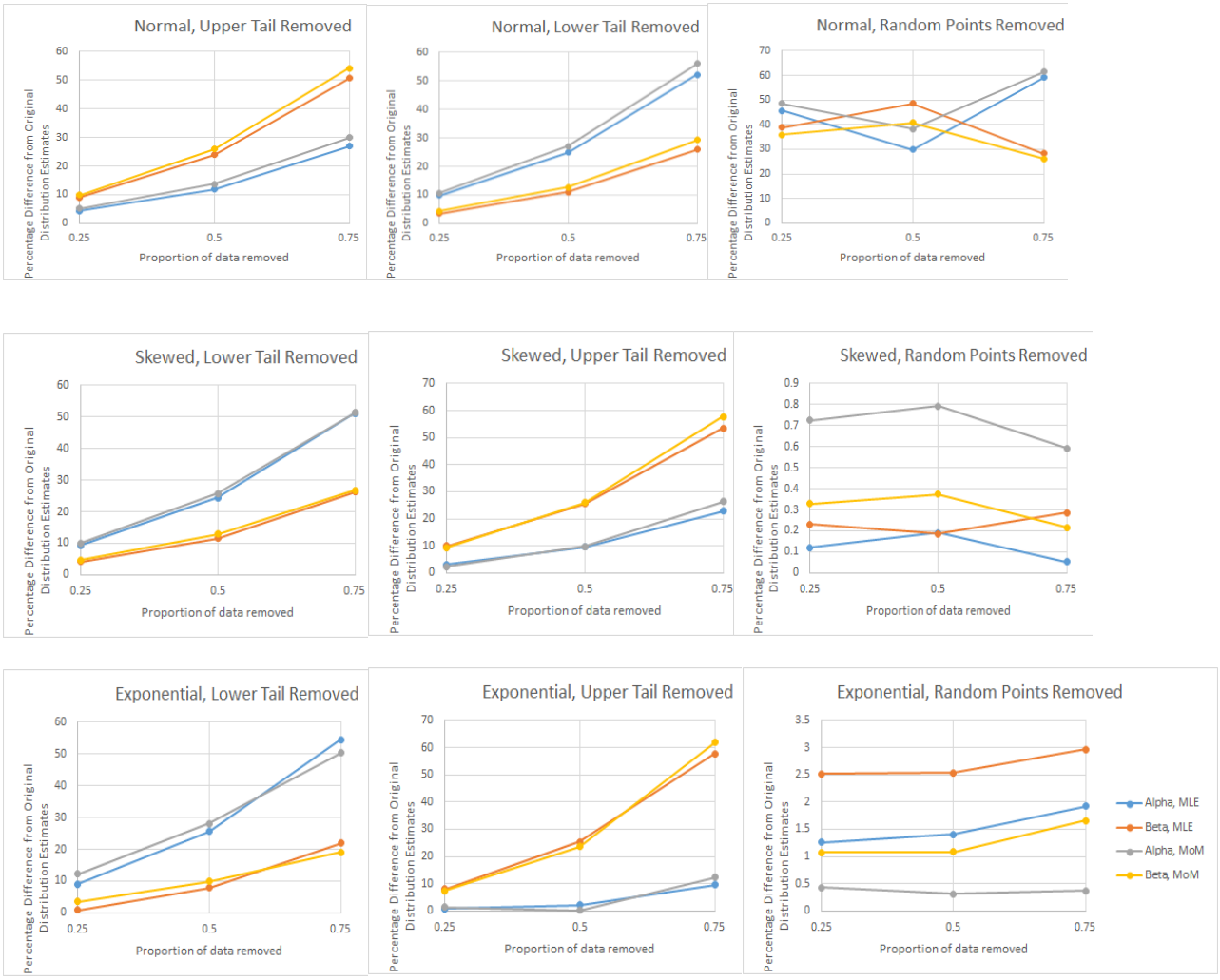

Figure 2. Scatterplots showing the performance of each estimation technique for every case.

The x-axis of each graph is the proportion of data that was removed, while the y-axis indicates the percentage difference between the original parameters of a given distribution and the estimates we found for our data with removed observations.

It is evident from Figure 2 that for all distributions, when data are systematically removed from tails, the difference between our original parameters and estimates increases as the percentage of removed data increases. However, this difference does not show any clear pattern when data is removed randomly.

Moreover, the difference is statistically significant at the $\alpha<0.001$ threshold in every case involving the upper and lower tail being removed. However, when data was removed randomly, for every

distribution, the difference was statistically significant at $\alpha<0.001$ only in the case where 75% of the data was removed from exponential distributions. In all other cases, it was not significant even at $\alpha<0.1$.

Finally, with the normal and skewed data, the difference between our actual parameter and the estimate was greater for the Method of Moments estimates that the Maximum Likelihood estimates. However, for exponential data, the opposite was true.

Discussion and Conclusion

As discussed above, Figure 2 and our hypothesis tests show that the differences are always significant within every distribution when data is removed from either tail. Hence, we can conclude that for any of the three distributions, when at least 25% of the data is missing from either one of the tails, our observed estimate will be significantly different from the actual parameters of our distributions. Moreover, as the percentage of missing data increases, this difference also increases.

On the other hand, we found that no matter how much data was randomly removed from the normal and slightly skewed distributions, our estimates were not significantly different from the original parameter values. Hence, when data is missing randomly throughout datasets with such distributions, our results suggest that our observed estimate will not be significantly different from the actual parameters of our distributions.

Therefore, random missing data does not have a great impact on relatively slightly skewed and normally distributed data. However, in the case of the exponential-shaped distribution, when at least 75% of the data are removed randomly, the difference between estimates and the actual value becomes significant.

Finally, as it can be seen from Figure 2, there is no obvious evidence suggesting which estimation technique is better between Method of Moments and Maximum Likelihood Estimation. However, Maximum Likelihood Estimation only performs slightly better in most of the cases.

Suggestions for Further Research

We only removed data randomly and from the tails of three different distributions. We therefore suggest adding more variety to the process of removing data points and testing the estimates on more types of distributions. With more computational time, we also hope to look into increasing the sample sizes of each distribution, which will increase the accuracy of the results.