Estimation of Population Parameters Using Sample Extremes from Nonconstant Sample Sizes

Alexander Bruno
Valparaiso University, alexander.bruno@valpo.edu

Follow this and additional works at: https://scholar.valpo.edu/sires

Recommended Citation
https://scholar.valpo.edu/sires/73

This Poster Presentation is brought to you for free and open access by the Office of Sponsored and Undergraduate Research at ValpoScholar. It has been accepted for inclusion in Summer Interdisciplinary Research Symposium by an authorized administrator of ValpoScholar. For more information, please contact a ValpoScholar staff member at scholar@valpo.edu.
Estimation of Population Parameters Using Sample Extremes from Nonconstant Sample Sizes

Alexander Bruno, Advisor: Dr. Tiffany Kolba

Abstract

We examine the accuracy and precision of parameter estimates for both the normal and exponential distribution when using only a collection of sample extremes. That is, we consider a collection of m random variables, where each of the m random variables is either the maximum or minimum of a sample of n_j independent, identically distributed random variables drawn from a normal or exponential distribution with unknown parameters. Previous work by Capaldi and Kolba (2019) derived estimators for the population parameters assuming the n_j sample sizes are constant. Since sample sizes are often not constant in applications, we utilize Mathlab to perform simulations to assess how the estimators from Capaldi and Kolba perform when the sample sizes are themselves random variables. Additionally, we explore how varying the mean, standard deviation, and probability distribution of the sample sizes affects the estimation error. Furthermore, we compare our results to those in the case where the sample sizes are drawn from a uniform distribution. Our estimation framework is applied to a biological example involving plant pollination.

Biological Application

The motivation for our research comes from a biological setting. We examine pollen tube lengths in two different sub-populations, Columbia (Col) and Landsberg (Ler), of the flowering plant. Arabidopsis thaliana. We wish to know the average pollen tube length in each sub-population. However, existing biological procedures can only measure the longest pollen tube in each plant. Using the maximum pollen tube lengths from a set of these flowers, we seek to estimate the population parameters which represents the mean pollen tube length.

Figure 1, obtained from Stawnor et al (2016), displays the pollen tube growth after 3 hours (image A), 6 hours (image B), and 9 hours (image C). The pollen tubes are dyed in blue, and the picture illustrates the difficulty that biologists would have in trying to measure the length of every pollen tube. However, the longest pollen tube is easily distinguishable and measurable from the images.

Figure 2

In the biological application, the pollen tube lengths were analyzed to follow an exponential distribution with unknown mean and standard deviation. In this section, we present the general framework for estimating μ using only a collection of sample extremes. We consider the case of sample maximums (which was used in the biological application) and the case of sample minimums separately.

Sample Maximum

Let X_{ij} \sim \text{Exponential}(\mu) for i = 1,\ldots,n and Y_j = \max(X_{ij}) for j = 1,\ldots,m. Set \hat{\mu} = \frac{1}{m} \sum_{j=1}^{m} Y_j. Previous work by Capaldi and Kolba proved that E(\hat{\mu}) = \mu and var(\hat{\mu}) = \frac{\mu^2}{m}. We consider the impact on the accuracy and precision of our estimation by varying different portions of the above formulation. We begin by looking at estimating \mu as either a uniform or normal random variable and consider the effect of increasing the range or standard deviation of \mu. The mean value of \mu was held constant at 1000. Upon examining our results in Figures 3 and 4, it is clear that we do not see a significant loss of accuracy or precision as the range or standard deviation of \mu increases.

Sample Minimum

We also created simulations for varying the values of \mu and \sigma. Here we set \mu to be uniformly distributed with a mean of 1000 and a range of 20. From Figures 5 and 6, we observe that as \sigma increases, the precision of the estimator increases. For \mu, the precision decreases as \mu increases. For both, the accuracy remains roughly constant, with a mean estimation error very close to zero. These results are analogous to the behavior when \mu is constant.

 concludes simulations for varying the values of \mu and \sigma. Here we set \mu to be uniformly distributed with a mean of 1000 and a range of 20. From Figures 5 and 6, we observe that as \sigma increases, the precision of the estimator increases. For \mu, the precision decreases as \mu increases. For both, the accuracy remains roughly constant, with a mean estimation error very close to zero. These results are analogous to the behavior when \mu is constant.

Conclusion

In addition to the explorations described above, we also derived a new unbiased estimator for the mean \mu from an exponential distribution when the sample sizes \{n_j\} follow a uniform distribution. We consider the case where X_{ij} \sim \text{Normal}(\mu, \sigma^2) for i = 1,\ldots,n and j = 1,\ldots,m, where \mu and \sigma^2 are unknown population parameters. While a normal distribution is not a good fit for the pollen tube lengths, which must be positive, a normal distribution is often a good population model in many applications. We again seek to estimate the unknown population parameters using only the sample extremes X_{ij} = \max(X_{ij}) and Y_j = \min(X_{ij}). Due to the symmetry of the normal distribution, estimation with the sample maximums or sample minimums is exactly analogous, so we consider here only the case of sample maximums. The recommended estimators from Capaldi and Kolba are \hat{\mu} = \frac{Y}{\bar{X}} = \frac{Y}{\bar{X}} - \mu_0 and \hat{\sigma}^2 = \frac{S^2}{\bar{X}} - \sigma^2, where Y is the mean of the sample maximums, \bar{X} is the standard deviation of the sample maximums, and \mu_0 and \sigma_0 are constants that depend only upon \sigma. The goal of this project was to assess the performance of the estimators when \mu is not constant for each sample. Figures 8 and 9 illustrate the estimation error of the mean \mu and variance \sigma^2, respectively, when \mu is drawn from a Normal distribution with mean 1000 and varying standard deviations. The accuracy and precision of the estimation is not significantly affected by the standard deviation of \mu. Although the estimators tend to overestimate \mu and underestimate \sigma, this behavior is seen even in the case where \mu is constant (standard deviation is zero). Figure 10 compares the estimation error between Normal, Uniform, and Poisson distributions for \mu with the same standard deviation. Again, the accuracy and precision is not affected by the type of distribution for \mu.

References

Capaldi A, Kolba TN. Using the sample maximum to estimate the parameters for the underlying distribution. PLOS ONE. 2015 Apr;10(4):e0123043.