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 minimum or maximum of a sample of n_i 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_i 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 examine exact distributional results 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 (Lei), 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 in a sub-population which represents the mean pollen tube length.

In the biological application, the pollen tube lengths were assumed to follow an exponential distribution with unknown mean \( \mu \) and \( \sigma \). In this section, we present the general framework for estimating \( \mu \) 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 \( X_{i} \sim \text{Exponential}(\mu) \) 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 \( \hat{\mu} \) is an unbiased estimator of \( \mu \). 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 modeling in such a uniform or normal random variable and consider the effect of increasing the range of standard deviation of n. The mean value of n 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 n increases.

We also considered simulations for varying the values of \( \mu \) and \( n \). Here we set \( n \) to be uniformly distributed with a mean of 1000 and a range of 20. From Figures 5 and 6, we can observe that as \( n \) increases, the precision of the estimation increases. For \( \mu \), the precision decreases as the variance 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 \( n \) is constant.

Sample Minimum

We now consider \( W_j = \min(X_{ij}) \), for \( j = 1, \ldots, m \), where \( X_{ij} \) is the sample of m random variables, where each of the m random variables is either the minimum or maximum of a sample of n_i independent, identically distributed random variables drawn from a normal or exponential distribution with unknown parameters. Previous work by Capaldi and Kolba proved that \( \hat{\mu} = Y - \frac{1}{m} \sum_{j=1}^{m} Y_j \) and \( \hat{\sigma} = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (Y_j^2 - \hat{\mu}^2)} \). Previous work by Capaldi and Kolba proved that \( \hat{\mu} \) is an unbiased estimator for \( \mu \). We now 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 \( Y_j = \max(X_{ij}) \) and \( W_j = \min(X_{ij}) \). 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 \( \hat{\mu} = Y - \frac{1}{m} \sum_{j=1}^{m} Y_j \) and \( \hat{\sigma} = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (Y_j^2 - \hat{\mu}^2)} \), where \( Y_j \) is the mean of the sample maximums, \( X_{ij} \) is the standard deviation of the sampled maximums, and \( k_0 \) and \( c_m \) are constants that depend only upon n. The goal of this project was to assess the performance of the estimators when \( n \) is not constant for each sample. Figures 8 and 9 illustrate the estimation error of the mean and variance \( \sigma^2 \), respectively, when \( n \) is drawn from a Normal distribution with mean 1000 and varying standard deviations. The accuracy and precision of the estimators is not significantly affected by the standard deviation of \( n \). Although the estimators tend to overestimate \( \mu \) and underestimate \( \sigma^2 \), this behavior is seen even in the case where \( n \) is constant (standard deviation is zero). Figure 10 compares the estimation error between Normal, Uniform, and Poisson distributions for \( n \) with the same standard deviation. Again, the accuracy and precision is not affected by the type of distribution for \( n \).

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 \( n \) follow a uniform distribution. However, this new estimator did not show any improvement in the accuracy or precision of the estimation compared to the estimator from Capaldi and Kolba, which assumed \( n \) is constant. Overall, we conclude that the estimation framework from Capaldi and Kolba performs well in the case where the sample sizes \( n \) widely vary. This result is useful for researchers who wish to estimate unknown population parameters using sample extremes since they do not need to worry about modeling the distribution of the sample sizes and can simply use the average sample size value for \( n \).

References