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 Matlab 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 derive new unbiased estimators 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 parameter μ which represents the mean pollen tube length.

Figure 1, obtained from Swanson 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.

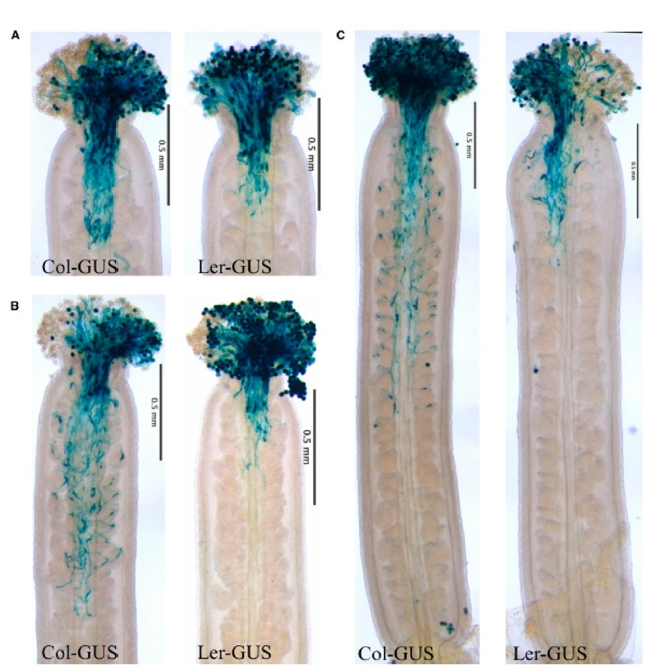

Figure 1

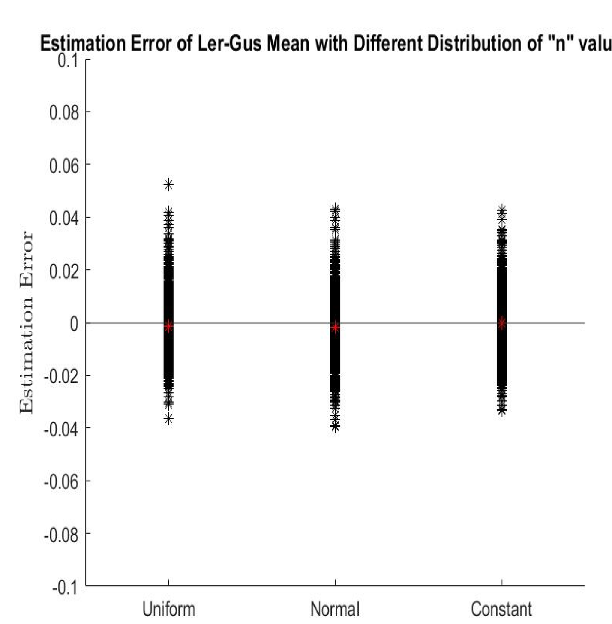

Figure 2

Previous work from Capaldi and Kolba (2019) estimated the mean pollen tube length assuming the number of pollen tubes was the same in every plant. However, in actuality not all plants will have the same number of pollen tubes. The total amount of pollen placed on each plant varies with values of 1040 ± 236. Furthermore, not all pollen that is placed on a plant germinates and produces a pollen tube. The total amount of germinated pollen varies not only from plant to plant, but also by the type of sub-population of plant. Using data from Swanson et al (2016), we consider both a uniform distribution and a normal distribution for the total germinated pollen (n) in both sub-populations:

- Germinated Ler-Gus pollen is *uniform* over interval (342, 1112)
  Germinated Col-Gus pollen is *uniform* over interval (640, 1226)

- Germinated Ler-Gus pollen is *normal* with mean= 727 and sd= 279
  Germinated Col-Gus pollen is *normal* with mean= 933 and sd= 243

In order to assess how the estimation framework from Capaldi and Kolba performs in the case of nonconstant sample sizes for the number of pollen tubes in each plant, we chose a reasonable guess for μ to be 0.2 mm. A specific value for μ was needed in order to check the error of the estimation. We then used Matlab to simulate values for the number of pollen tubes using a uniform distribution and normal distribution and plotted the estimation error ($\hat{\mu} - \mu$). Figure 2 illustrates that all three cases of n uniformly distributed, normally distributed, or constant provided roughly the same level of accuracy and precision. We ran 1000 simulations for each case and plotted the estimation error (black dots) as well as the mean estimation error for each distribution (red dot).

Exponential Distribution

In the biological application, the pollen tube lengths were assumed to follow an exponential distribution with unknown mean μ. 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 $X_{ij} \overset{iid}{\sim}$ Exponential(μ) for $i = 1, ..., n$ and $Y_j = \max\{X_{ij}\}_{i=1}^{n}$ for $j = 1, ..., m$. Set $\hat{\mu} = \frac{\bar{Y}}{H_n} = \frac{1}{mH_n}\sum_{j=1}^{m} Y_j$. Previous work by Capaldi and Kolba proved that $E(\hat{\mu}) = \mu$ and $var(\hat{\mu}) = \frac{\mu^2 G_n}{mH_n^2}$ where $H_n = \sum_{i=1}^{n}\frac{1}{i}$, $G_n = \sum_{i=1}^{n}\frac{1}{i^2}$. We consider the impact on the accuracy and precision of our estimation by varying different portions of the above formulation. We began by looking at modeling n as either a uniform or normal random variable and consider the effect of increasing the range or 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 are able to conclude from this that nonconstant sample sizes do not significantly hinder the performance of our estimations and that it is reasonable to use the estimator from Capaldi and Kolba even in the case of widely varying sample sizes.

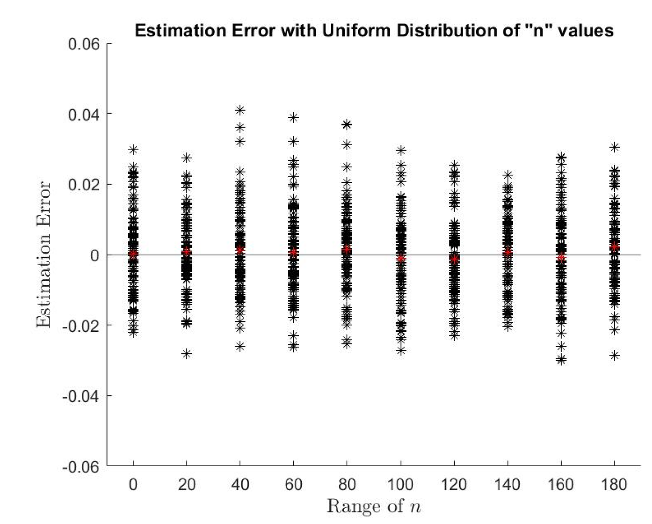

Figure 3

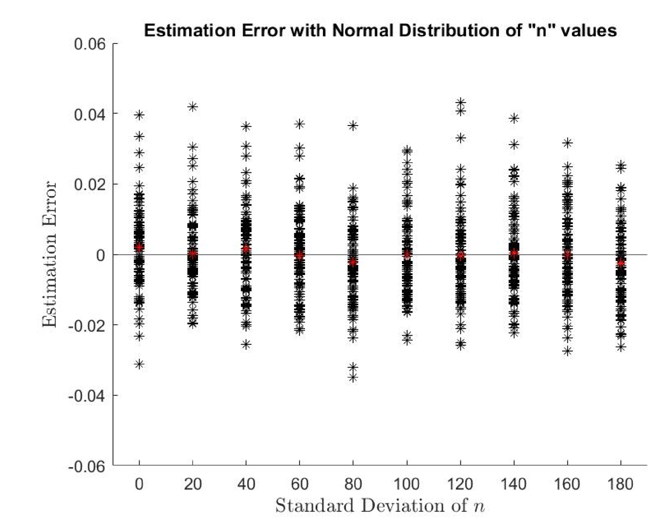

Figure 4

We also created simulations for varying the values of μ and m. Here we set n to be uniformly distributed with a mean of 1000 and a range of 20. From Figures 5 and 6 below, we can observe that as m increases, the precision of the estimation increases. For μ, the precision decreases as the value 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.

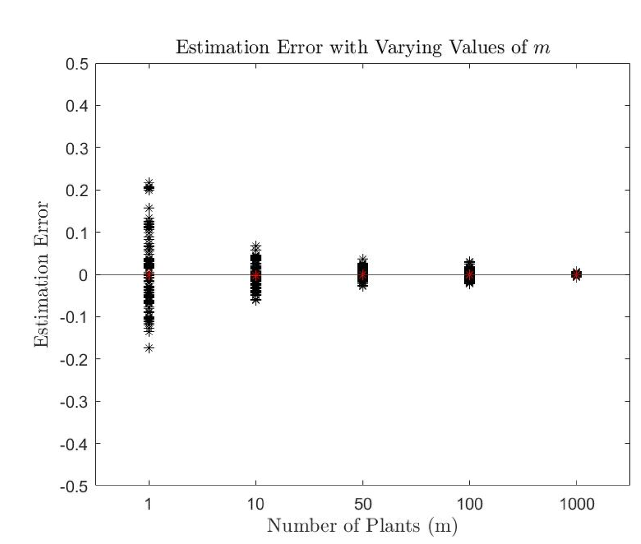

Figure 5

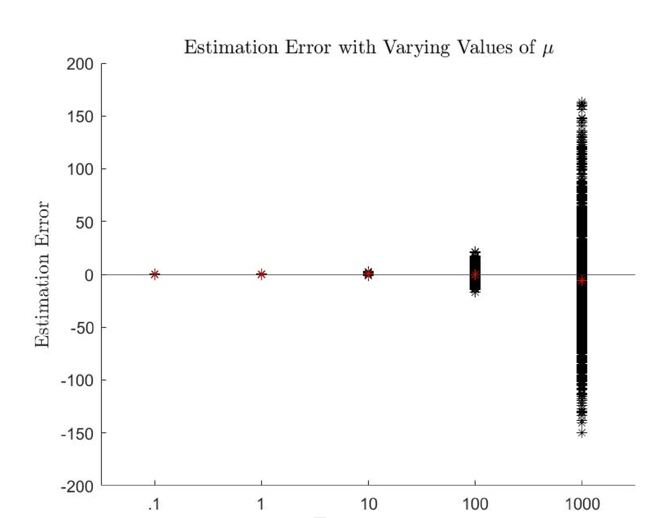

Figure 6

Sample Minimum

We now consider $W_j = \min\{X_{ij}\}_{i=1}^{n}$, where X_{ij} is defined the same as above. It directly follows from the work of Capaldi and Kolba that $\hat{\mu} = n\bar{W} = \frac{n}{m}\sum_{j=1}^{m} W_j$ is an unbiased estimator for μ when n is constant. Here, we are multiplying by n to adjust for the inherent underestimation of the mean that comes from using minimums. This is the opposite of what we do with the sample maximums, which would lend themselves to an overestimation, thus requiring a division adjustment. Again, we consider drawing the n values from different distributions to assess the performance when n is not constant. Figure 7 compares drawing n from Normal, Uniform, and Poisson distributions with the same standard deviation. As with the case of the sample maximum, the distribution of n does not significantly affect the results, and the estimation error is similar to the case where n is constant.

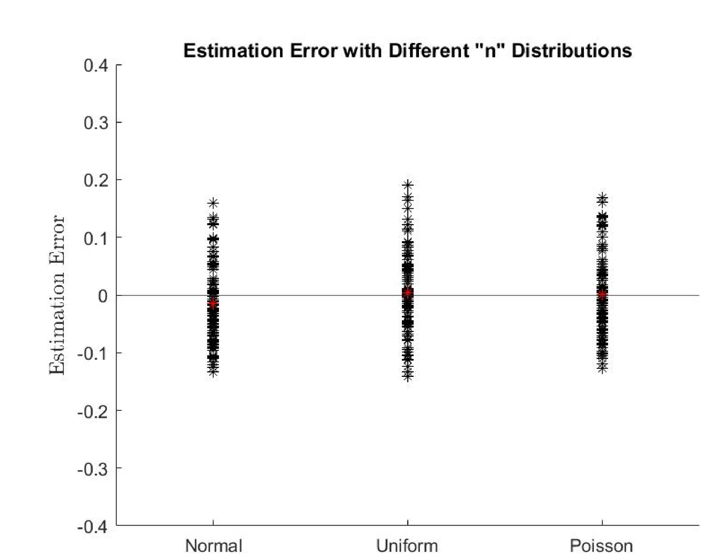

Figure 7

Normal Distribution

We now consider the case where $X_{ij} \overset{iid}{\sim}$ Normal(μ, σ^2) for $i = 1, ..., n$ and $j = 1, ..., m$, where μ and σ^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 $Y_j = \max\{X_{ij}\}_{i=1}^{n}$ and $W_j = \min\{X_{ij}\}_{i=1}^{n}$. 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 $\hat{\mu} = \bar{Y} - \frac{k_n}{\sqrt{c_n}}S_Y$, and $\hat{\sigma}^2 = \frac{S_Y^2}{c_n}$, where $\bar{Y}$ is the mean of the sample maximums, S_Y is the standard deviation of the sample maximums, and k_n and c_n 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 μ and variance σ^2, respectively, when n 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 n. Although the estimators tend to overestimate μ and underestimate σ^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.

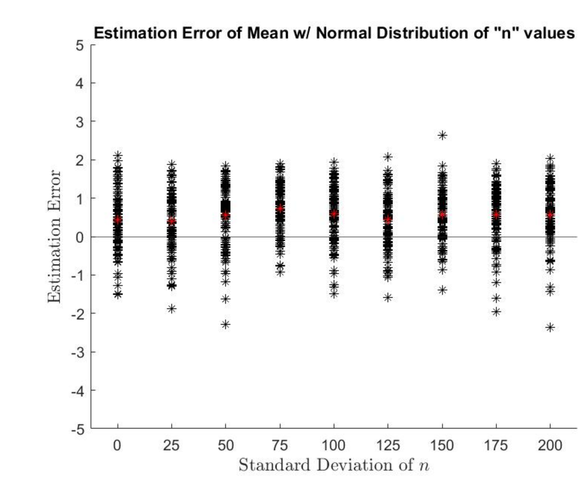

Figure 8

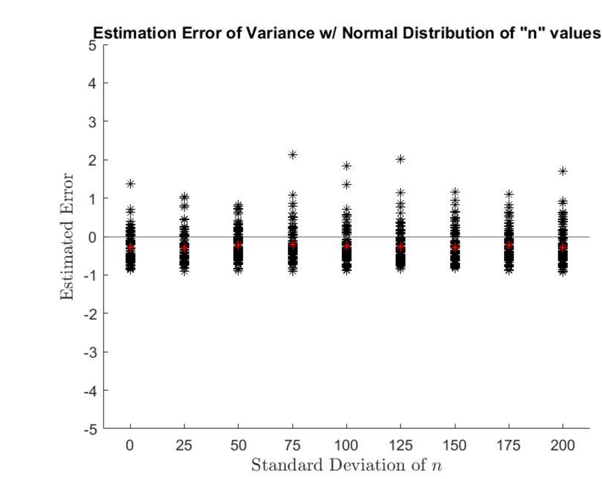

Figure 9

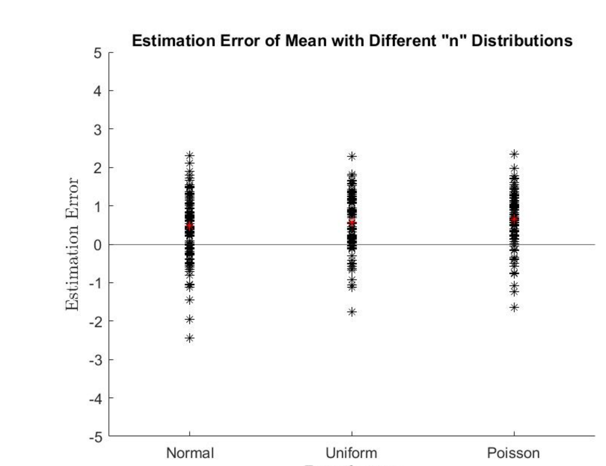

Figure 10

Conclusion

In addition to the explorations described above, we also derived a new unbiased estimator for the mean μ 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 even 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

- Capaldi A, Kolba TN. Using the sample maximum to estimate the parameters for the underlying distribution. PLOS ONE. 2019 Apr; 14(4):1-9.
- Swanson RJ, Hammond AT, Carlson AL, Gong H, Donovan TK. Pollen performance traits reveal prezygotic nonrandom mating and interference competition in *Arabidopsis thaliana*. American Journal of Botany. 2016 Feb; 103(3):498-513.