Summer 7-23-2021

Optimizing Percentile Matching

Joshua Megchelsen
Olivet Nazarene University, jrmegchelsen@olivet.edu

Optimizing Percentile Matching

"Using an Alternative Technique for Estimating Population Parameters in the Exponential Distribution"

A Pence-Boyce Research Project by

Student: Joshua Megchelsen
Faculty: Dr. Dale Hathaway

JULY 23, 2021

Abstract

Point estimation is a technique used in statistics to estimate unknown parameters in populations of data by using samples of data from that population. Percentile matching is a method of point estimation which selects one piece of data from a sample and assumes that the percentile that piece of data represents in the sample is equal to that same percentile's theoretical value in the population. That assumption is then used to project what the unknown parameter is. The fundamental question our research sought to answer was which percentile should be matched from the sample to the population to produce the best point estimates for the exponential distribution. By evaluating percentiles that naturally occurred in samples, we were able use order statistics to calculate the variance and expected value of error for point estimates created by different percentiles. We concluded that creating point estimates using the 80th percentile will consistently be optimal.

Table of Contents

1. Introduction:

Point estimation is a technique used in statistics that uses sample data to estimate population parameters that are unknown. There are multiple approaches to point estimation, including most prominently, the method of moments and the method of maximum likelihood. However, there are situations when these more common methods are not ideal or even possible. An alternative in these situations is a newer, lesser-known approach to point estimation called percentile matching. Our research concentrated on understanding and optimizing percentile matching for the exponential distribution and comparing it to the other point estimation techniques.

1.1 Statistical Distributions

Discrete distributions and continuous densities are used to describe how a population of data is distributed. Mathematically, continuous random variables are represented with a probability density function (PDF) which describes the theoretical distribution of the population. This is the PDF of the exponential distribution:

$$f(x;\theta) = \tfrac{1}{\theta} e^{-\frac{x}{\theta}} \text{ for all } x > 0$$

It is a function of x, which is an individual piece of data, given θ, which is the parameter that will describe a particular population. If θ is given, then one can enter any value for x in the domain to find the density for that value in the population. Integrating this function over the defined domain will result in 1 as all the probabilities for any distribution added together will sum to 1. Consider this for the exponential distribution:

$$\int_0^\infty \frac{1}{\theta} e^{-\frac{x}{\theta}}\,dx = \lim_{a\to\infty}\int_0^a \frac{1}{\theta} e^{-\frac{x}{\theta}}\,dx = \lim_{a\to\infty}\left(1 - e^{\frac{-a}{\theta}}\right) - \left(1 - e^{\frac{0}{\theta}}\right) = 1 - 0 = 1$$

Integrating a PDF on the interval $[0, x]$ will result in a cumulative distribution function (CDF). This function will describe the probability of an individual piece of data being less than or equal to a given value x in a distribution. The CDF of the exponential distribution is:

$$F(x) = \int_0^x \frac{1}{\theta} e^{-\frac{x}{\theta}}dx = 1 - e^{\frac{-x}{\theta}} = P(X \le x)$$

What makes each population's distribution unique are the parameters that are associated with the population. The parameters will not only determine the PDF's and CDF's of distributions, but they also define important elements of the distribution such as the mean and variance. In the exponential distribution (which only has one parameter θ)

$$\mu = \theta;\ \sigma^2 = \theta^2$$

where μ is the mean of the population and $\sigma^2$is the variance of the population. If the parameters are unknown for a population, they can be estimated using sample data from that population.

1.2 Point Estimation for Exponential Distribution

Point estimators are rules, usually expressed as functions, that describe how an estimation for a population parameter can be made. They take information from a sample to infer information about a population. Ideal point estimators are unbiased. Mathematically this means that the expected value of the point estimate is equal to the parameter it is estimating. Point estimates can be judged on other factors such as efficiency, consistency, and sufficiency. The best point estimate for a population is considered the minimum variance unbiased estimator (MVUE).

For the exponential distribution, the MVUE for θ is $\bar{x}$, which is the mean of the sample. It is then by definition, an unbiased and sufficient estimator. Since the variance of $\bar{x}$ is $V(\bar{x}) = \frac{1}{n^2}\theta$, $\bar{x}$ is a consistent estimator as $\lim_{n\to\infty} V(\bar{x}) = \lim_{n\to\infty} \frac{1}{n^2}\theta = 0$. $\bar{x}$ is also the estimator produced by both the method of moments and the method of maximum likeliness. In our study of percentile matching, we will then compare point estimates for θ created by percentiles to $\bar{x}$.

1.3 Introduction to Percentile Matching

The theory of percentile matching is as follows: The theoretical percentile p of a population is expressed as a function in terms of a parameter θ and is set equal to the smoothed empirical percentile $\hat{p}$ from the sample of data. θ is then solved for in this equation and the resulting $\hat{\theta}$ is considered the point estimate for the parameter θ. In other words, we assume that the value of a selected percentile from a sample is equal to the value of that theoretical percentile in the population and then can deduce what the parameter must then be. The theoretical percentile p is solved for by setting the CDF of the distribution equal to the selected percentile p, (where $0 \leq p \leq 1$), and the variable x is solved for. This is done because $\forall p,\ p = P(x \leq p) = 1 - e^{\frac{-x}{\theta}} = F(x)$. In the exponential distribution:

$$F(x) = p$$

$$1 - e^{\frac{-x}{\theta}} = p$$

$$x = -\theta \ln(1 - p)$$

The value for x will then be set equal to the smoothed empirical percentile $\hat{p}$ which is:

$$\hat{p} = (1-h)x_j + hx_{j+1}$$

where,

$$j = \lfloor (n+1)p \rfloor, \qquad h = (n+1)p - j,$$

and x_j is equal to the j^{th}piece of data in the ordered sample. Essentially, $\hat{p}$ is an appropriately weighted number to represent the p^{th} empirical percentile of any sample. So,

$$\hat{p} = x = -\theta \ln(1-p)$$

$$\frac{-\hat{p}}{\ln(1-p)} = \hat{\theta} \text{ or } -\frac{x}{\ln(1-p)} = \hat{\theta}$$

$\hat{\theta}$ is the point estimator, p is the percentile that is selected to match, and $\hat{p} = x$ is the value from the sample that corresponds to the selected percentile.

Consider an example from the exponential distribution. Using the statistical programming language R, the following dataset with $n = 10$ was simulated from a population with an exponential distribution where $\theta = 10$. The data has been ordered:

0.8952618, 1.4665267, 1.5653028, 4.0474807, 6.6689763,

10.7436686, 13.1427591, 15.1162930, 17.3070972, 18.6535244

Letting $p = .5$ for the 50th percentile

$$\hat{p} = (1-h)x_j + hx_{j+1} = (1-.5)6.6689763 + .5(10.7436686) = 8.706322$$

Using the CDF of the exponential and equating it to p:

$$1 - e^{\frac{-x}{\theta}} = .5 \Rightarrow \hat{p} = x = -\theta \ln(1-.5)$$

Now we can set the sample 50th percentile equal to the theoretical 50th percentile and solve for the point estimator of θ:

$$8.706322 = -\theta \ln(1-.5) \Rightarrow \hat{\theta} = 12.56057$$

Recall that the parameter for this population was $\theta = 10$ and the point estimate using the percentile matching technique is $\hat{\theta} = 12.56057$. However, the results for the point estimate vary depending on what percentile is used. Table 1.3.1 shows the point estimators if nine evenly spaced percentiles for this sample were used.

Table 1.3.1

Percentile Used	Smoothed Percentile from the Sample	Resulting Point Estimate
0.1	0.95238829	9.039328292
0.2	1.48628192	6.660653697
0.3	2.30995617	6.476362327
0.4	5.09607894	9.976161537
0.5	8.70632245	12.56056822
0.6	12.1831229	13.29613241
0.7	14.52423283	12.06358879
0.8	16.86893636	10.48125947
0.9	18.51888168	8.042648125

The results from these calculations show that the 40th percentile followed by the 80th and 10th percentile produced the point estimates that were closest to the original parameter of 10. However, this is not always the case with other samples.

If a different sample of simulated data (shown below in order) from the same population was used still with $n = 10$:

6.203843, 6.737784, 7.451190, 7.965972, 10.822362

12.433454, 14.452704, 15.903059, 17.034323, 44.919423

then the resulting output of the percentile calculations are:

Table 2.3.2

Percentile Used	Smoothed Percentile from the Sample	Resulting Point Estimate
0.1	6.25724	59.3888
0.2	6.88047	30.8343
0.3	7.60562	21.3237
0.4	9.10853	17.831
0.5	11.6279	16.7755
0.6	13.645	14.8916
0.7	15.468	12.8474
0.8	16.8081	10.4434
0.9	42.3109	18.2972

The results from this sample are very different from the first sample. This is seen in the range of point estimates created in conjunction with the percentiles that created the best point estimates. Finding which percentiles in a sample will produce the best point estimates consistently is how we will optimize this method of point estimation.

2. Order Statistics

2.1 Dealing with Percentiles.

In our research, we always calculated the percentiles of a sample as $\frac{k}{n+1}$, where k is the rank of the statistic in the sorted sample of size n. Therefore, $0 < k \leq n$ and $0 < \frac{k}{n+1} < 1$, with $k, n \in \mathbb{Z}$. While there are other methods of computing the percentile of a statistic, this method was selected for its consistency for percentile matching (Schoonjans). Any sample will produce n different percentiles that can be matched with their theoretical values in the population. If one wants to match a percentile that is not in the sample provided, the percentile must be estimated with the smoothed empirical percentile. However, when this smoothed empirical percentile is calculated, it weighs the values from the sample in a linear method and does not consider the distribution from which the sample came. For example, in the exponential distribution, a sample of size 9 would provide the 70th and 80th percentile with the 7th and 8th pieces of data when the sample was sorted. The weighted estimation for the 75th percentile would be exactly halfway in between the values for the 7th and 8th pieces of data. However, the exponential distribution will always have a higher density for a lower percentile. The PDF for the distribution is strictly decreasing in the given domain (the function is monotonic) for all θ values. Therefore, any percentile p will have a smaller distance between the percentile $p - a$ than the percentile $p + a$ where $0 < p - a < p < p + a < 1$. So, this implies that the actual 75th percentile of a sample would fall closer to the 70th percentile than the 80th. This error will change in proportion to how large θ is.

A second reason why estimating percentiles creates error in point estimation is that it assumes that two percentiles are at their expected value instead of just one. Even if one estimated the value of a percentile that did not exist in a sample with respect to the population from which it came, one would have to assume that both the percentile above and below it was at their expected value. Percentile matching with a percentile that is already in the sample only assumes that one percentile is at its expected value. Essentially weighting percentiles with respect to the population is creating a weighted average from two different point estimates of the percentiles closest to the estimated one. It will be shown later that for any sample there is an order statistic that will produce a minimum amount of error for a point estimate. Weighting this statistic with any other statistic to

produce a point estimate from an estimated percentile will necessarily raise the expected error from the minimum.

For this reason, we did not estimate any percentiles; we only used percentiles in the form $\frac{k}{n+1}$ so as to not lose precision when matching sample percentiles with theoretical percentiles. Since we are only focused on matching percentiles that naturally fall on order statistics, this more specific and precise form of percentile matching could be called order statistics matching.

2.2 Order Statistics Calculations

To find which order statistic would produce the best estimate we wanted to know how statistics were distributed in an exponential sample. This will allow us to see the expected value and variance for order statistics with a given n. Knowing the expected value and variance will allow us to create unbiased point estimators and see which statistics are more consistently closer to those expected values. It is well known that given a sample from an exponential distribution of size n with $x_1 < x_2 < \cdots < x_k < \cdots < x_n$ the PDF for x_k will be,

$$f(x,n,k) = \frac{n!}{(k-1)!\,(n-k)!}\left(\left(1-e^{-\frac{x}{\theta}}\right)^{k-1}\left(\frac{1}{\theta}e^{-\frac{x}{\theta}}\right)\left(1-\left(1-e^{-\frac{x}{\theta}}\right)\right)^{n-k}\right)$$

$$= \frac{n!}{(k-1)!\,(n-k)!}\left(\left(1-e^{-\frac{x}{\theta}}\right)^{k-1}\left(\frac{1}{\theta}e^{-\frac{x}{\theta}}\right)\left(e^{-\frac{x}{\theta}}\right)^{n-k}\right)$$

$$= \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\left(-e^{-\frac{x}{\theta}}+1\right)^{k-1}\left(e^{-\frac{(n-k+1)x}{\theta}}\right)\right)$$

Using the properties of binomial expansion (the last term will always be positive, but all other terms will alternate signs):

$$= \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\left(\pm\binom{k-1}{0}e^{-\frac{(k-1)x}{\theta}} \pm \binom{k-1}{1}e^{-\frac{(k-2)x}{\theta}} \pm \cdots + \binom{k-1}{k-1}e^{-\frac{(k-k)x}{\theta}}\right)\left(e^{-\frac{(n-k+1)x}{\theta}}\right)\right)$$

$$= \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\left(\pm\binom{k-1}{0}e^{-\frac{nx}{\theta}} \pm \binom{k-1}{1}e^{-\frac{(n-1)x}{\theta}} \pm \cdots + \binom{k-1}{k-1}e^{-\frac{(n-k+1)x}{\theta}}\right)\right)$$

$$= \frac{n!}{\theta(k-1)!\;(n-k)!}\left(\left(\pm\frac{(k-1)!\;e^{-\frac{nx}{\theta}}}{0!\,(k-1)!} \pm \frac{(k-1)!\;e^{-\frac{(n-1)x}{\theta}}}{1!\,(k-2)!} \pm \cdots + \frac{(k-1)!\;e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!}\right)\right)$$

Equation 2.2.1:

$$f(n,k,x) = \frac{n!}{\theta(n-k)!}\left(\left(\pm\frac{e^{-\frac{nx}{\theta}}}{0!\,(k-1)!} \pm \frac{e^{-\frac{(n-1)x}{\theta}}}{1!\,(k-2)!} \pm \cdots + \frac{e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!}\right)\right)$$

When $\mathrm{n} = 9, \theta = 1$ the distributions of the 9 statistics in the sample look like this:

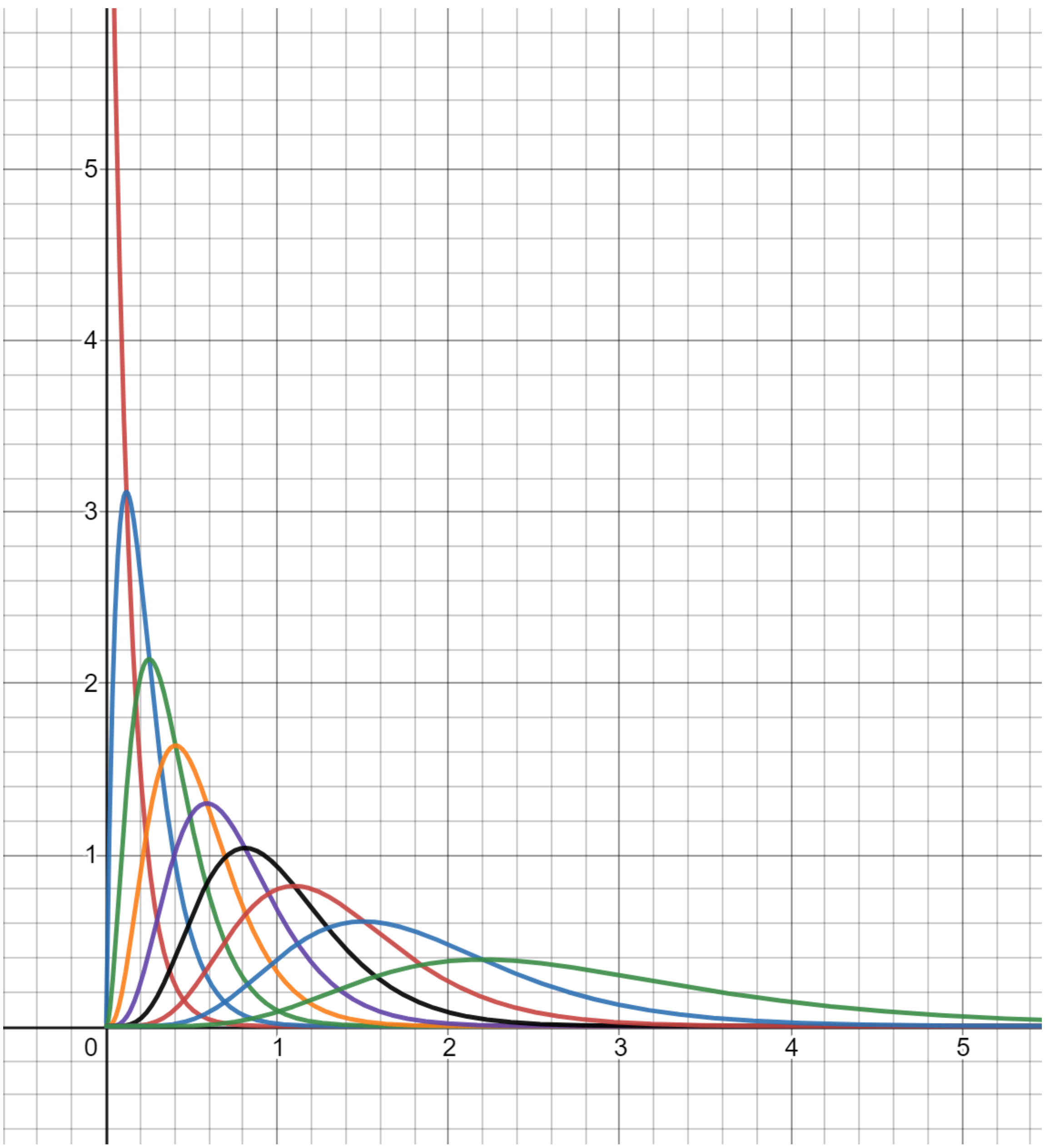


Figure 2.2.2

Expected value and variance can be found using:

$$E(x) = \int_0^\infty x \cdot PDF\, dx$$

$$E(x^2) = \int_0^\infty x^2 \cdot PDF\, dx$$

$$V(x) = E(x^2) - [E(x)]^2$$

We will solve for the expected value as functions of n, k and the notation will be $\mathrm{E}_{\mathrm{n,k}}(\mathrm{x})$.

Finding $\mathrm{E}_{\mathrm{n,k}}(\mathrm{x})$:

$$E_{n,k}(x) = \int_0^\infty x \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\left(\pm\binom{k-1}{0}e^{-\frac{nx}{\theta}} \pm \binom{k-1}{1}e^{-\frac{(n-1)x}{\theta}} \pm \cdots + \binom{k-1}{k-1}e^{-\frac{(n-k+1)x}{\theta}}\right)\right)dx$$

$$= \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\pm\binom{k-1}{0}\int_0^\infty x\, e^{-\frac{nx}{\theta}}dx \pm \binom{k-1}{1}\int_0^\infty x\, e^{-\frac{(n-1)x}{\theta}}dx \pm \cdots + \binom{k-1}{k-1}\int_0^\infty x\, e^{-\frac{(n-k+1)x}{\theta}}dx\right)$$

Notice:

$$\int_0^\infty x e^{\frac{-cx}{\theta}}\, dx = \lim_{b\to\infty}\int_0^{\mathrm{b}} x e^{\frac{-cx}{\theta}}\, dx = \lim_{b\to\infty} -\frac{\theta(cx+\theta)e^{\frac{-cx}{\theta}}}{c^2}\Bigg|_0^b$$

$$\lim_{b\to\infty} -\frac{\theta(cb+\theta)e^{\frac{-cb}{\theta}}}{c^2} + \frac{\theta(0+\theta)e^{\frac{0}{\theta}}}{c^2} = \frac{\theta^2}{c^2}$$

So therefore,

$$\frac{n!}{\theta(k-1)!\,(n-k)!}\left(\pm\binom{k-1}{0}\int_0^\infty x\, e^{-\frac{nx}{\theta}}dx \pm \binom{k-1}{1}\int_0^\infty x\, e^{-\frac{(n-1)x}{\theta}}dx \pm \cdots + \binom{k-1}{k-1}\int_0^\infty x\, e^{-\frac{(n-k+1)x}{\theta}}dx\right)$$

$$= \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\pm\frac{\binom{k-1}{0}\theta^2}{n^2} \pm \frac{\binom{k-1}{1}\theta^2}{(n-1)^2} \pm \cdots + \frac{\binom{k-1}{k-1}\theta^2}{(n-k+1)^2}\right)$$

$$= \frac{n!\,\theta}{(k-1)!\,(n-k)!}\left(\pm\frac{\frac{(k-1)!}{0!\,(k-1)!}}{n^2} \pm \frac{\frac{(k-1)!}{1!\,(k-2)!}}{(n-1)^2} \pm \cdots + \frac{\frac{(k-1)!}{(k-1)!\,0!}}{(n-k+1)^2}\right)$$

$$= \frac{n!\,\theta}{(n-k)!}\left(\pm\frac{1}{0!\,(k-1)!\,n^2} \pm \frac{1}{1!\,(k-2)!\,(n-1)^2} \pm \cdots + \frac{1}{(k-1)!\,0!\,(n-k+1)^2}\right)$$

<u>Equation 2.2.3:</u>

$$E_{n,k}(x) = \theta \frac{n!}{(n-k)!} \sum_{i=1}^{k} \frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}$$

Finding $E_{n,k}(x^2)$:

$$E_{n,k}(x^2) = \int_0^\infty x^2 \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\left(\pm\binom{k-1}{0}e^{-\frac{nx}{\theta}} \pm \binom{k-1}{1}e^{-\frac{(n-1)x}{\theta}} \pm \cdots + \binom{k-1}{k-1}e^{-\frac{(n-k+1)x}{\theta}}\right)\right)dx$$

$$= \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\pm\binom{k-1}{0}\int_0^\infty x^2\, e^{-\frac{nx}{\theta}}dx \pm \binom{k-1}{1}\int_0^\infty x^2\, e^{-\frac{(n-1)x}{\theta}}dx \pm \cdots + \binom{k-1}{k-1}\int_0^\infty x^2\, e^{-\frac{(n-k+1)x}{\theta}}dx\right)$$

Notice:

$$\int_0^\infty x^2 e^{\frac{-cx}{\theta}}\,dx = \lim_{b\to\infty}\int_0^{\mathrm{b}} x^2 e^{\frac{-cx}{\theta}}\,dx = \lim_{b\to\infty} -\left.\frac{\theta(c^2x^2 + 2\theta cx + 2\theta^2)e^{\frac{-cx}{\theta}}}{c^3}\right|_0^b$$

$$\lim_{b\to\infty} -\frac{\theta(c^2b^2 + 2\theta cb + 2\theta^2)e^{\frac{-cb}{\theta}}}{c^3} - \frac{\theta(0+0+2\theta^2)e^{\frac{0}{\theta}}}{c^3} = \frac{2\theta^3}{c^3}$$

So therefore,

$$\frac{n!}{\theta(k-1)!\,(n-k)!}\left(\pm\binom{k-1}{0}\int_0^\infty x^2\, e^{-\frac{nx}{\theta}}dx \pm \binom{k-1}{1}\int_0^\infty x^2\, e^{-\frac{(n-1)x}{\theta}}dx \pm \cdots + \binom{k-1}{k-1}\int_0^\infty x^2\, e^{-\frac{(n-k+1)x}{\theta}}dx\right)$$

$$= \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\pm\frac{\binom{k-1}{0}2\theta^3}{n^3} \pm \frac{\binom{k-1}{1}2\theta^3}{(n-1)^3} \pm \cdots + \frac{\binom{k-1}{k-1}2\theta^3}{(n-k+1)^3}\right)$$

$$= \frac{n!\,\theta^2}{(k-1)!\,(n-k)!}\left(\pm\frac{\frac{2(k-1)!}{0!\,(k-1)!}}{n^3} \pm \frac{\frac{2(k-1)!}{1!\,(k-2)!}}{(n-1)^3} \pm \cdots + \frac{\frac{2(k-1)!}{(k-1)!\,0!}}{(n-k+1)^3}\right)$$

$$= \frac{n!\,\theta^2}{(n-k)!}\left(\pm\frac{2}{0!\,(k-1)!\,n^3} \pm \frac{2}{1!\,(k-2)!\,(n-1)^3} \pm \cdots + \frac{2}{(k-1)!\,0!\,(n-k+1)^3}\right)$$

<u>Equation 2.2.4</u>:

$$E_{n,k}(x^2) = \theta^2 \frac{n!}{(n-k)!} \sum_{i=1}^{k} \frac{2(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^3}$$

Finding $V_{n,k}(x)$:

$$V_{n,k}(x) = E_{n,k}(x^2) - \left[E_{n,k}(x)\right]^2$$

Equation 2.2.5:

$$\boxed{V_{n,k}(x) = \theta^2\left(\frac{n!}{(n-k)!}\sum_{i=1}^{k}\frac{2(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^3} - \left(\frac{n!}{(n-k)!}\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}\right)^2\right)}$$

It appears that lower order statistics will always have lower variability:

$$V_{n,k}(x) < V_{n,k+1}(x), \forall k < n.$$

We suspect this both because the exponential distribution will have higher density for lower order statistics (see figure 2.2.2) and because equation 2.2.5 indicates this when multiple k values are tested.

3. Point Estimation with Percentile Matching

Setting the CDF of the exponential distribution equal to a percentile will determine the equation of a point estimate given that percentile.

$$F(x) = p$$

$$1 - e^{\frac{-x}{\theta}} = \frac{k}{n+1}$$

$$e^{\frac{-x}{\theta}} = 1 - \frac{k}{n+1}$$

$$\hat{\theta} = \frac{-x}{\ln\left(1 - \frac{k}{n+1}\right)}$$

Finding the expected value for any point estimate will then be simple as $\frac{-1}{\ln\left(1-\frac{k}{n+1}\right)}$ is a constant:

$$E_{n,k}\left(\frac{-x}{\ln\left(1 - \frac{k}{n+1}\right)}\right) = \frac{-1}{\ln\left(1 - \frac{k}{n+1}\right)}\, E_{n,k}(x)$$

Equation 3.0.1:

$$E_{n,k}\left(\frac{-x}{\ln\left(1-\frac{k}{n+1}\right)}\right)=\theta\left(\left(\frac{-1}{\ln\left(1-\frac{k}{n+1}\right)}\right)\left(\frac{n!}{(n-k)!}\right)\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}\right)$$

3.1 Finding an Unbiased Point Estimator

For any given n, the point estimate is not unbiased as the expected value does not equal θ. However, as $n \to \infty$, $\left(\frac{n!}{(n-k)!}\right)\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!(i-1)!(n-k+i)^2} \to -\ln\left(1-\frac{k}{n+1}\right)$. (Since it will approach this value from above, all point estimates will actually be a small amount over θ (Appendix A). This happens because the expected value of a percentile in a sample will approach the theoretical value of that percentile in a population. When the CDF of a population is set equal to a percentile, the value of x that is solved for will be the value of that percentile in the distribution.

$$F(x)=p$$

$$x=-\theta\ln\left(1-\frac{k}{n+1}\right)$$

So, as $n \to \infty$, the expected value of a percentile will approach $-\theta\ln\left(1-\frac{k}{n+1}\right)$. Therefore as $n \to \infty\ \forall k$, the point estimate is unbiased.

Proposition 3.1.1:

When $n \to \infty$, $E\left(-\frac{x}{\ln\left(1-\frac{k}{n+1}\right)}\right)=\theta\left(\frac{-1}{\ln\left(1-\frac{k}{n+1}\right)}\right)\left(-\ln\left(1-\frac{k}{n+1}\right)\right)=1\theta$

Therefore, as $n \to \infty$, any percentile produces an unbiased estimator.

However, by altering the point estimator, we can make the point estimate unbiased for any value of n. This can be done by multiplying a point estimate by the constant $\frac{-\ln\left(1-\frac{k}{n+1}\right)(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!(i-1)!(n-k+i)^2}}\right)$.

$$E\left(\left(-\frac{x}{\ln(1-p)}\right)\frac{-\ln\left(1-\frac{k}{n+1}\right)(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\right)=$$

$$\int_0^\infty\left(-\frac{x}{\ln(1-p)}\right)\frac{-\ln(1-p)\,(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\left(\frac{n!}{\theta(n-k)!}\right)\left(\frac{e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!}-\frac{e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!}\pm\cdots\pm\frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!}\right)dx$$

$$=\frac{(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\int_0^\infty x\left(\frac{n!}{\theta(n-k)!}\right)\left(\frac{e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!}-\frac{e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!}\pm\cdots\pm\frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!}\right)dx$$

$$=\frac{(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)E_{n,k}(x)$$

$$=\frac{(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\left(\left(\frac{n!}{(n-k)!}\right)\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}\right)\theta$$

$$=1\theta$$

<u>Proposition 3.1.2:</u>

$$\boxed{\therefore\ \frac{x(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!(i-1)!(n-k+i)^2}}\right)\text{ is an unbiased estimator of }\theta\ \forall\ n,k}$$

While this is an unbiased estimator, some percentiles will still produce better point estimators than others. The output of the unbiased point estimator will be distributed differently for different values of n and k. All of these distributions will have a mean of θ but they will have different variances and different expected errors. Less variance is better as it indicates that the point estimate's distribution is less spread out around θ. An error of a point estimate can be calculated by taking the absolute value of the difference of θ and the point estimate. The expected value of an error can be determined by taking the average of all the errors in all possible point estimates weighted appropriately to how likely each error is to occur.

3.2 Variance in the Distribution of Point Estimates from Percentiles

To find the variance of an unbiased estimator ($\hat{x}$) we will need its expected value and the expected value of the estimator squared.

$$V_{n,k}(\hat{x})=E_{n,k}(\hat{x}^2)-\left(E_{n,k}(\hat{x})\right)^2$$

Since it is unbiased its expected value will be θ.

$$E_{n,k}(\hat{x}) = 1\theta$$

$$V_{n,k}(\hat{x}) = E_{n,k}(\hat{x}^2) - (1\theta)^2$$

$$E_{n,k}(\hat{x}^2) = \int_0^\infty \left(\frac{x(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\right)^2 \left(\frac{n!}{\theta(n-k)!}\right)\left(\frac{e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!} \pm \frac{e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!} \pm \cdots + \frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!}\right) dx$$

$$= \left(\frac{(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\right)^2 \int_0^\infty x^2 \frac{n!}{\theta(k-1)!\,(n-k)!}\left(\left(\pm\binom{k-1}{0}e^{-\frac{nx}{\theta}} \pm \binom{k-1}{1}e^{-\frac{(n-1)x}{\theta}} \pm \cdots + \binom{k-1}{k-1}e^{-\frac{(n-k+1)x}{\theta}}\right)\right) dx$$

$$= \left(\frac{(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\right)^2 E_{n,k}(x^2)$$

$$= \left(\frac{(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\right)^2 \theta^2 \frac{n!}{(n-k)!}\sum_{i=1}^{k}\frac{2(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^3}$$

<u>Equation 3.2.1</u>:

$$\boxed{V_{n,k}(\hat{x}) = \theta^2\left(\left(\frac{(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\right)^2 \frac{n!}{(n-k)!}\sum_{i=1}^{k}\frac{2(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^3} - 1\right)}$$

This variance can be compared to the variance of the MVUE of the exponential distribution $\bar{x}$:

$$V(\bar{x}) = \frac{1}{n}\theta^2$$

It appears that $V_{n,k}(\hat{x}) \to \frac{1}{k}$. This trend can be seen on the table in appendix B.

<u>Conjecture 3.2.2:</u>

$$\boxed{As\ n \to \infty\ \ \forall k, \qquad V_{n,k}(\hat{x}) \to \frac{1}{k}}$$

Since the variance of this percentile matched point estimate is approaching $\frac{1}{k}$ and

$k \leq n,\ V(\bar{x}) \leq V_{n,k}(\hat{x})$. The only time the variances will be equal is when $n = 1$.

Because $V_{n,k}(\hat{x}) \to \frac{1}{k}$, it would seem that the highest k value possible would then make for the minimum variance for a point estimate, thus indicating that the highest order statistic would be optimal for point estimation. However, since variance is approaching $\frac{1}{k}$ and not equal to it, that is not the case. Lower k values are closer to $\frac{1}{k}$ for any given n value because lower k values start approaching $\frac{1}{k}$ sooner. For example, $V_{n,2}(\hat{x})$ will approach $\frac{1}{2}$ for all $n \geq 2$, where as $V_{n,100}(\hat{x})$ will only approach $\frac{1}{100}$ after $n \geq 100$. So for any n value, $V_{n,2}(\hat{x})$ is closer to $\frac{1}{2}$ than $V_{n,100}(\hat{x})$ is to $\frac{1}{100}$. Higher k's having lower variances, but lower k's being closer to $\frac{1}{k}$ results in a balance where the minimum variance occurs when $\frac{k}{n+1}$ is closer to .8, indicating that the optimal percentiles for point estimation for the exponential distribution is the 80th. A complete table of calculations for variance for all $k \leq n < 30$ can be found in appendix B.

Conjecture 3.2.3:

$$\boxed{As\ n \to \infty, \qquad \min\left(V_{n,k}(\hat{x})\right) \to \frac{1}{\lfloor .8(n+1) \rfloor}}$$

For a point estimate to be considered consistent, its variance must approach 0 as $n \to \infty$. Since the variance of a percentile point estimate is a function of both n and k, and appears to be approach $\frac{1}{k}$, it is difficult to say if it is a consistent estimator. Certainly, it is not consistent $\forall k$, but we will consider point estimates created by sufficiently large k as pseudo-consistent estimators because the highest values of $k \to \infty$ as $n \to \infty$.

Table 3.2.4 indicates which order statistic (and its respective percentile) produced the minimum variance as a point estimator for $n \leq 30$. It also indicates what the order statistic should be multiplied by in order to obtain the unbiased point estimator. Finally, the table shows how much variance there is as a factor of θ^2, as well as the variance of the MVUE for that n.

Table 3.2.4

N	Order Stat	Percentile Used	Point Estimate	Variance	Variance of MVUE
1	1	0.5	1x	1	1.000
2	2	0.667	.667x	0.556	0.500
3	3	0.75	.545x	0.405	0.333
4	4	0.8	.480x	0.328	0.250
5	5	0.833	.438x	0.281	0.200
6	5	0.714	.690x	0.234	0.167
7	6	0.75	.627x	0.202	0.143
8	7	0.778	.582x	0.179	0.125
9	8	0.8	.547x	0.161	0.111
10	8	0.727	.700x	0.147	0.100
11	9	0.75	.658x	0.133	0.091
12	10	0.769	.624x	0.123	0.083
13	11	0.7857	.595x	0.114	0.077
14	12	0.8	.571x	0.106	0.071
15	12	0.75	.673x	0.099	0.067
16	13	0.765	.646x	0.09323	0.063
17	14	0.777	.624x	0.0879	0.059
18	15	0.789	.603x	0.08321	0.056
19	16	0.8	.583x	0.0791	0.053
20	16	0.762	.660x	0.075	0.050
21	17	0.773	.640x	0.07165	0.048
22	18	0.783	.621x	0.0685	0.045
23	19	0.792	.605x	0.0656	0.043
24	19	0.8	.591x	0.0631	0.042
25	20	0.769	.653x	0.0605	0.040
26	21	0.778	.636x	0.0582	0.038
27	22	0.786	.621x	0.0561	0.037
28	23	0.793	.609x	0.05412	0.036
29	24	0.8	.596x	0.05234	0.034
30	24	0.774	.648x	0.0506	0.033

Even for very large samples, such as those of size 49, 99, and 349, the percentile with minimum variance was very near the 80th. Figures 3.2.5, 3.2.6 and 3.2.7 plot the minimum variance points against their relative percentiles for these sample sizes.

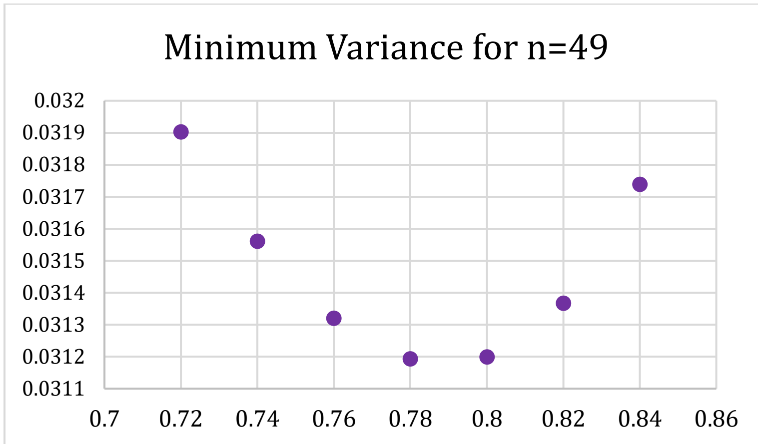


Figure 3.2.5

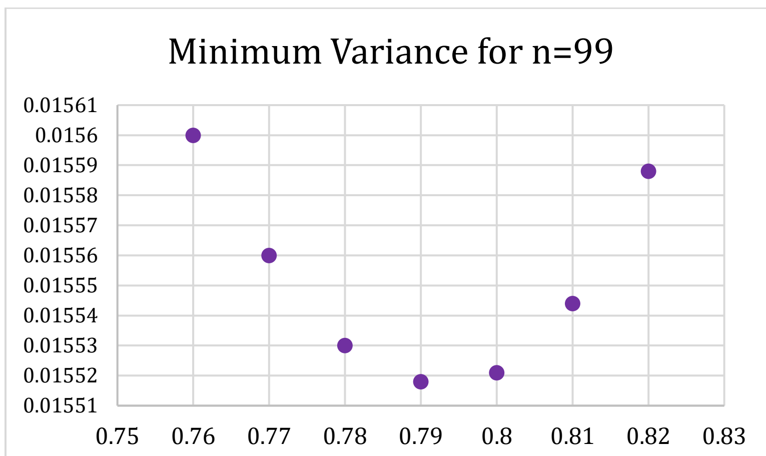


Figure 3.2.6

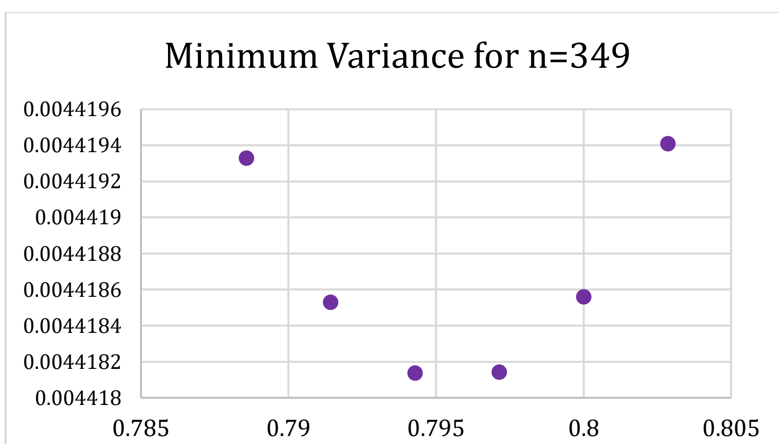


Figure 3.2.7

3.3 The Expected Value of an Error

Another way of evaluating how good a point estimate will be is to look at the expected value of an error. The error of an unbiased point estimator is defined as the distance between θ and the value of the point estimate for a given x value greater than 0:

<u>Equation 3.3.1:</u>

$$Error = \left|\theta - \frac{x}{\ln(1-p)}\left(\frac{\ln(1-p)\,(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!\,(i-1)!\,(n-k+i)^2}}\right)\right)\right|$$

This is a linear absolute value function with a y-intercept at θ, and a slope of $\pm\frac{(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!(i-1)!(n-k+i)^2}}\right)$. This function will be referred to as the error line of a percentile. As already shown,

$$n \to \infty,\ \left(\frac{n!}{(n-k)!}\right)\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!(i-1)!(n-k+i)^2} \to -\ln(1-p),$$

$$\therefore n \to \infty,\ \frac{(n-k)!}{n!}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-1)!(i-1)!(n-k+i)^2}}\right) \to \frac{1}{-\ln(1-p)}.$$

As the function will always be positive, the x-intercept is the expected value of $x$for that order statistic, which is the inverse of the slope. Table 3.3.2 and figure 3.3.3 show the error lines for the 9 percentiles graphed when $n = 9, \theta = 1$.

Table 3.3.2

Percentile (k)	Color	x-intercept	Slope
10th (1)	Red	.111	$\pm$9.00
20th (2)	Blue	.236	$\pm$4.24
30th (3)	Green	.379	$\pm$2.63
40th (4)	Orange	.546	$\pm$1.83
50th(5)	Purple	.746	$\pm$1.34
60th (6)	Black	.996	$\pm$1.00
70th (7)	Red	1.329	$\pm$.752
80th (8)	Blue	1.829	$\pm$.547
90th (9)	Green	2.829	$\pm$.353

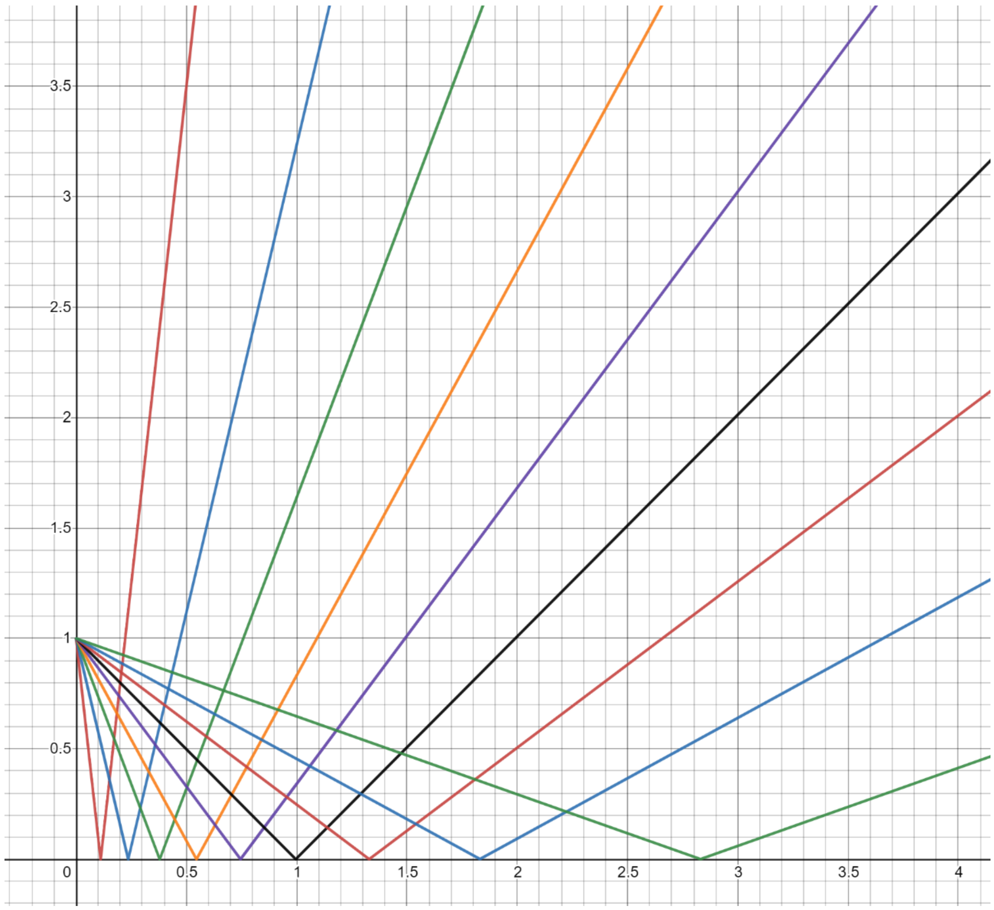


Figure 3.3.3

The height for any x value of each line is representative of the error of the point estimate when the value of the given percentile is that x value. For example, the purple line represents the 5th order statistic, or in this sample the 50th percentile. When this statistic is equal to about 1.5, the unbiased point estimate created from the 50th percentile will have an error of about 1θ (or since in this population $\theta = 1$ just 1). The expected value of an error can be found by summing up all possible errors times the probability that error will happen. The probability of an error can be found by using the PDF for that order statistic. Mathematically, given n, k the expected value of an error is:

$$\int_{0}^{\infty} Error * f(n, k, x) dx$$

Figure 3.3.4 shows the PDF of the 50th percentile of a sample of size 9 overlaid with the error line for the 50th percentile. Notice that when the density is the highest in the PDF, the error is near the lowest.

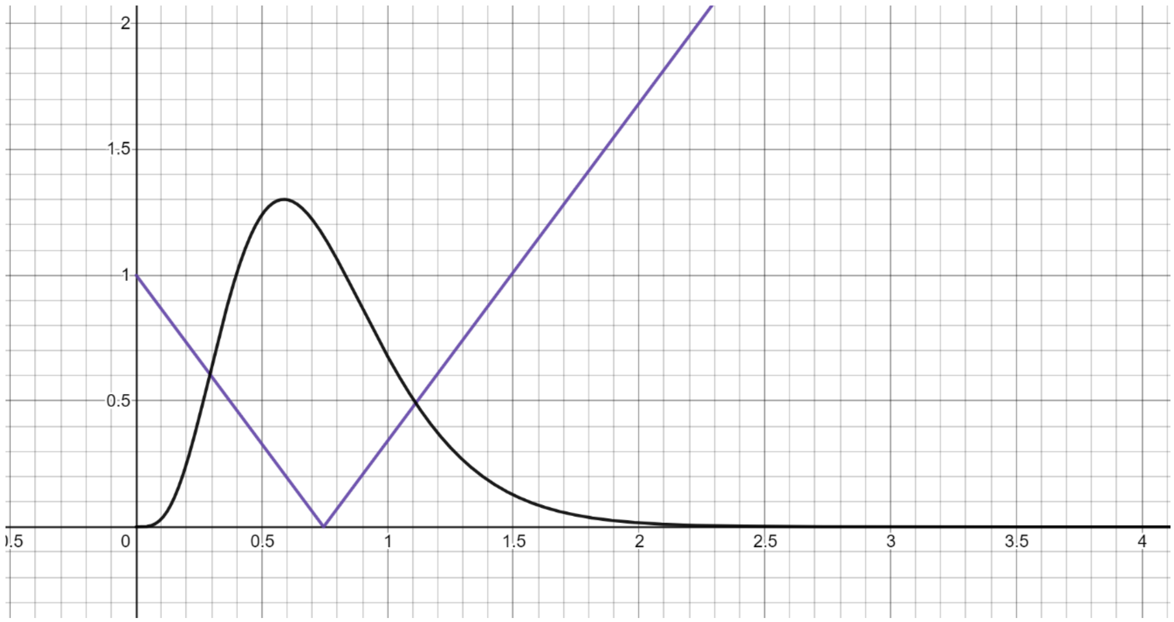


Figure 3.3.4

Figure 3.3.5 illustrates a geometric interpretation of the integral for expected error. The figure is the curve created when error line for the 50th percentile is multiplied by the PDF for the 50th percentile again with $n = 9, \theta = 1$:

Figure 3.3.5

$\int_0^\infty Error * f(9,5,x)dx$ is the calculation of the area under this curve. To get a sense of how different percentiles have different amounts of error, figure 3.3.6 shows this same graph with the addition of 10th (shown in red) and 80th (shown in green) percentiles in a sample of size 9:

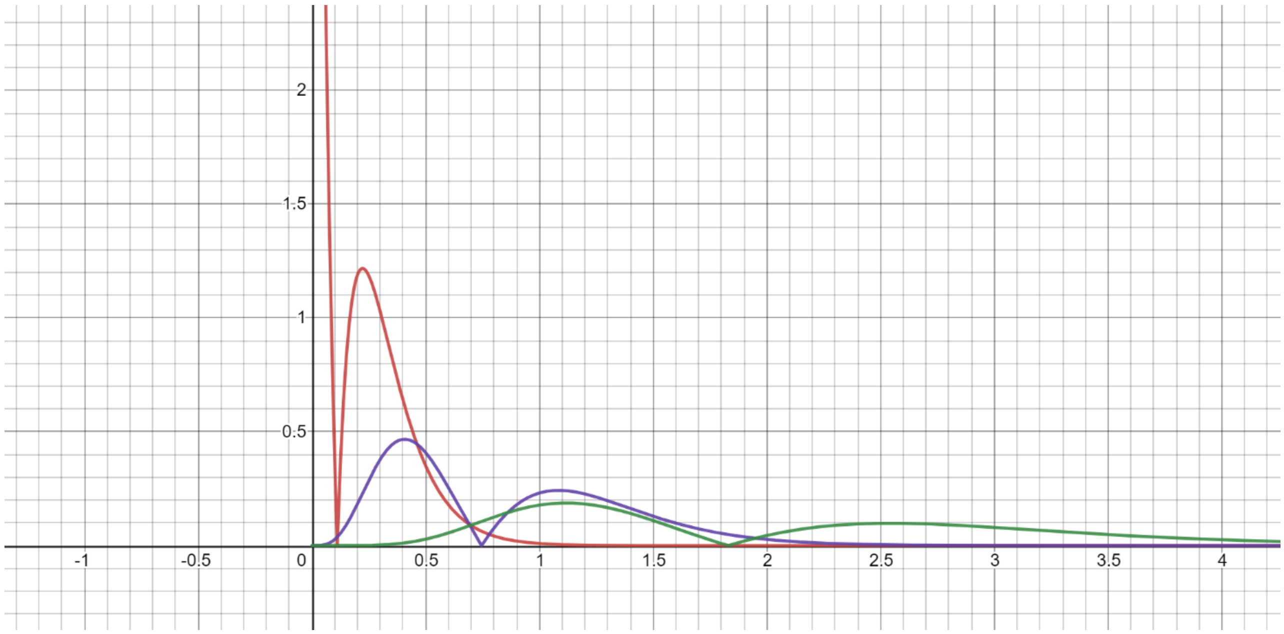


Figure 3.3.6

Clearly there is less area under the 80th percentile curve, indicating that there is less expected error for point estimates created by the 80th percentile than the 50th or the 10th. In order to evaluate what the expected error is for any n, k, we can evaluate the integral shown below:

$$E_{n,k}(Error) = \int_{0}^{\infty} Error * f(n,k,x)dx =$$

$$\int_0^\infty \left| \theta + \frac{x}{\ln(1-p)} \left(\frac{-\ln(1-p)\,(n-k)!}{n!} \left(\frac{1}{\sum_{i=1}^{k} \frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}} \right) \right) \right| \left(\frac{n!}{\theta(n-k)!} \right) \left(\frac{e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!} - \frac{e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!} \pm \cdots \pm \frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!} \right) dx$$

$$= \int_0^{E_{n,k}(x)} \left(\frac{n!}{(n-k)!} \right) \left(\frac{e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!} - \frac{e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!} \pm \cdots \pm \frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!} \right) dx$$

$$- \int_0^{E_{n,k}(x)} x \left(\frac{1}{\sum_{i=1}^{k} \frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}} \right) \left(\frac{1}{\theta} \right) \left(\frac{e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!} - \frac{e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!} \pm \cdots \pm \frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!} \right) dx$$

$$- \lim_{b\to\infty} \int_{E_{n,k}(x)}^{b} \left(\frac{n!}{(n-k)!} \right) \left(\frac{e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!} - \frac{e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!} \pm \cdots \pm \frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!} \right) dx$$

$$+ \lim_{b\to\infty} \int_{E_{n,k}(x)}^{b} x \left(\frac{1}{\sum_{i=1}^{k} \frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}} \right) \left(\frac{1}{\theta} \right) \left(\frac{e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!} - \frac{e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!} \pm \cdots \pm \frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!} \right) dx$$

$$= \frac{n!}{(n-k)!}\left(\frac{-\theta e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!\,(n-k+1)} + \frac{\theta e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!\,(n-k+2)} \pm \cdots \pm \frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!\,n}\right)\Bigg|_0^{E_{n,k}(x)}$$

$$-\frac{1}{\theta}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\right)\left(\frac{-\theta\big((n-k+1)x+\theta\big)e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!\,(n-k+1)^2} \pm \cdots \pm \frac{\theta\big((n)x+\theta\big)e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!\,(n)^2}\right)\Bigg|_0^{E_{n,k}(x)}$$

$$-\lim_{b\to\infty}\frac{n!}{(n-k)!}\left(\frac{-\theta e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!\,(n-k+1)} + \frac{\theta e^{-\frac{(n-k+2)x}{\theta}}}{(k-2)!\,1!\,(n-k+2)} \pm \cdots \pm \frac{e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!}\right)\Bigg|_{E_{n,k}(x)}^{b}$$

$$+\lim_{b\to\infty}\frac{1}{\theta}\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\right)\left(\frac{-\theta\big((n-k+1)x+\theta\big)e^{-\frac{(n-k+1)x}{\theta}}}{(k-1)!\,0!\,(n-k+1)^2} \pm \cdots \pm \frac{\theta\big((n)x+\theta\big)e^{-\frac{(n)x}{\theta}}}{0!\,(k-1)!\,(n)^2}\right)\Bigg|_{E_{n,k}(x)}^{b}$$

Recall that $E_{n,k}(x)$ is a function of n, k multiplied by θ (Equation 2.2.3). When evaluating the function then at $E_{n,k}(x)$, a θ can be factored out and the remaining coefficient will be referred to as v, such that $v = \left(\frac{n!}{(n-k)!}\right)\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!(i-1)!(n-k+i)^2}$

$$= \frac{n!}{(n-k)!}\left(\frac{-\theta e^{-(n-k+1)v}}{(k-i)!\,0!\,(n-k+1)} + \frac{\theta e^{-(n-k+2)v}}{(k-2)!\,1!\,(n-k+2)} \pm \cdots \pm \frac{\theta e^{-(n)v}}{0!\,(k-1)!\,n}\right)$$

$$-\frac{n!}{(n-k)!}\left(\frac{-\theta}{(k-i)!\,0!\,(n-k+1)} + \frac{\theta}{(k-2)!\,1!\,(n-k+2)} \pm \cdots \pm \frac{\theta}{0!\,(k-1)!\,n}\right)$$

$$-\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\right)\left(\frac{-\big((n-k+1)(\theta v+\theta\big)e^{-(n-k+1)v}}{(k-1)!\,0!\,(n-k+1)^2} \pm \cdots \pm \frac{\big((n)(-\theta v)+\theta\big)e^{-(n)v)}}{0!\,(k-1)!\,(n)^2}\right)$$

$$+\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\right)\left(\frac{-\theta}{(k-1)!\,0!\,(n-k+1)^2} \pm \cdots \pm \frac{\theta}{0!\,(k-1)!\,(n)^2}\right)$$

$$+\frac{n!}{(n-k)!}\left(\frac{-\theta e^{-(n-k+1)v}}{(k-1)!\,0!\,(n-k+1)} + \frac{\theta e^{-(n-k+2)v}}{(k-2)!\,1!\,(n-k+2)} \pm \cdots \pm \frac{\theta e^{-(n)v}}{0!\,(k-1)!\,n}\right)$$

$$-\left(\frac{1}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\right)\left(\frac{-\big((n-k+1)(-\theta v+\theta\big)e^{-(n-k+1)v}}{(k-1)!\,0!\,(n-k+1)^2} \pm \cdots \pm \frac{\big((n)(-\theta v)+\theta\big)e^{-(n)v}}{0!\,(k-1)!\,(n)^2}\right)$$

$$= \frac{\theta n!}{(n-k)!}\sum_{i=1}^{k}\frac{(-1)^{i-1}e^{-(n-k+1)v}}{(k-i)!\,(i-1)!\,(n-k+i)} - \frac{\theta n!}{(n-k)!}\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)}$$

$$- \frac{\theta}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\left(\sum_{i=1}^{k}\frac{(-1)^{i}\big((n-k+i)(v)+1\big)e^{-(n-k+1)v}}{(k-i)!\,(i-1)!\,(n-k+i)^2}\right)$$

$$- \frac{\theta}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\left(\sum_{i=1}^{k}\frac{(-1)^{i}}{(k-i)!\,(i-1)!\,(n-k+i)^2}\right) + \frac{\theta n!}{(n-k)!}\sum_{i=1}^{k}\frac{(-1)^{i-1}e^{-(n-k+i)v}}{(k-i)!\,(i-1)!\,(n-k+i)}$$

$$- \frac{\theta}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\left(\sum_{i=1}^{k}\frac{(-1)^{i}\big((n-k+i)(v)+1\big)e^{-(n-k+i)v}}{(k-i)!\,(i-1)!\,(n-k+i)^2}\right)$$

$$= \theta\left(\frac{2n!}{(n-k)!}\sum_{i=1}^{k}\frac{(-1)^{i}e^{(n-k+i)v}}{(k-i)!\,(i-1)!\,(n-k+i)} - 1 - \frac{2}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\sum_{i=1}^{k}\frac{(-1)^{i}\big((n-k+i)(v)+1\big)e^{-(n-k+i)v}}{(k-i)!\,(i-1)!\,(n-k+i)^2} + 1\right)$$

So, the equation for the expected value of an error given n, k:

<u>Equation 3.3.7:</u>

$$\boxed{= \theta\left(\frac{2n!}{(n-k)!}\sum_{i=1}^{k}\frac{(-1)^{i}e^{-(n-k+i)v}}{(k-i)!\,(i-1)!\,(n-k+i)} - \frac{2}{\sum_{i=1}^{k}\frac{(-1)^{i-1}}{(k-i)!\,(i-1)!\,(n-k+i)^2}}\sum_{i=1}^{k}\frac{(-1)^{i}\big((n-k+i)(v)+1\big)e^{-(n-k+i)v}}{(k-i)!\,(i-1)!\,(n-k+i)^2}\right)}$$

In this formula, θ is able to be factored out, meaning that the expected value of error can be written as a coefficient of θ. Similar to variance, regardless of what the value of θ is the expected value of error for any n, k will be in the same proportion to θ. Another way of thinking of this coefficient is it represents what percent of θ a given error will be. In the example where the 50th percentile in a sample with $n = 9, \theta = 1$ was equal to 1.5 and the resulting error was 1 or 100% of θ. A complete table of calculations for the expected value of error for all $k \leq n < 30$ can be found in appendix C.

Table 3.3.3 indicates which order statistic (and its respective percentile) produced the minimum expected error as a point estimator for all $n \leq 30$. The table indicates what the value of the order statistic should be multiplied by in order to get the unbiased point estimator. Finally, table 3.3.7 also shows how much expected error there is as a factor of θ.

Table 3.3.7

N	Order Stat	Percentile Used	Point Estimate	Expected Value of Error
1	1	0.5	1x	0.693147
2	2	0.667	.667x	0.517491
3	3	0.75	.545x	0.439652
4	4	0.8	.480x	0.39406
5	5	0.833	.438x	0.36343
6	5	0.714	.690x	0.341096
7	6	0.75	.627x	0.323899
8	7	0.778	.582x	0.310134
9	8	0.8	.547x	0.296647
10	9	0.818	.519x	0.283694
11	9	0.75	.658x	0.272811
12	10	0.769	.624x	0.263503
13	11	0.786	.595x	0.255426
14	12	0.8	.571x	0.246948
15	13	0.813	.549x	0.239455
16	13	0.765	.646x	0.232812
17	14	0.777	.624x	0.22687
18	15	0.789	.603x	0.221406
19	16	0.8	.583x	0.215872
20	17	0.81	.566x	0.210854
21	17	0.773	.640x	0.206277
22	18	0.783	.621x	0.202081
23	19	0.792	.605x	0.198102
24	20	0.8	.591x	0.194146
25	20	0.769	.653x	0.190489
26	21	0.778	.636x	0.187094
27	22	0.785	.621x	0.183933
28	23	0.793	.609x	0.18088
29	24	0.8	.596x	0.177873
30	24	0.774	.648x	0.175057

Again, the optimal percentile appears to be approaching the 80th percentile as n increases. Given the complexity and large numbers that are needed to calculate the expected value of error, it was difficult to calculate minimum error for very large samples accurately even with a computer algebra system. However, the minimum expected error for samples of size 49 and 74 were still at the 80th percentile. Figures 3.3.8 and 3.3.9 plot the minimum expected values of error against their respective percentiles for these sample sizes.

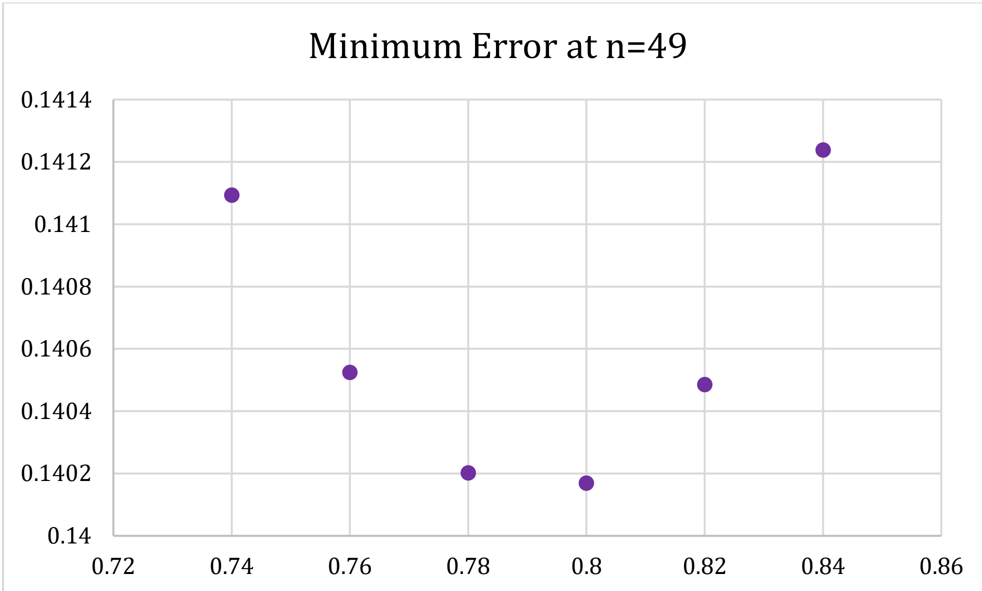


Figure 3.3.8

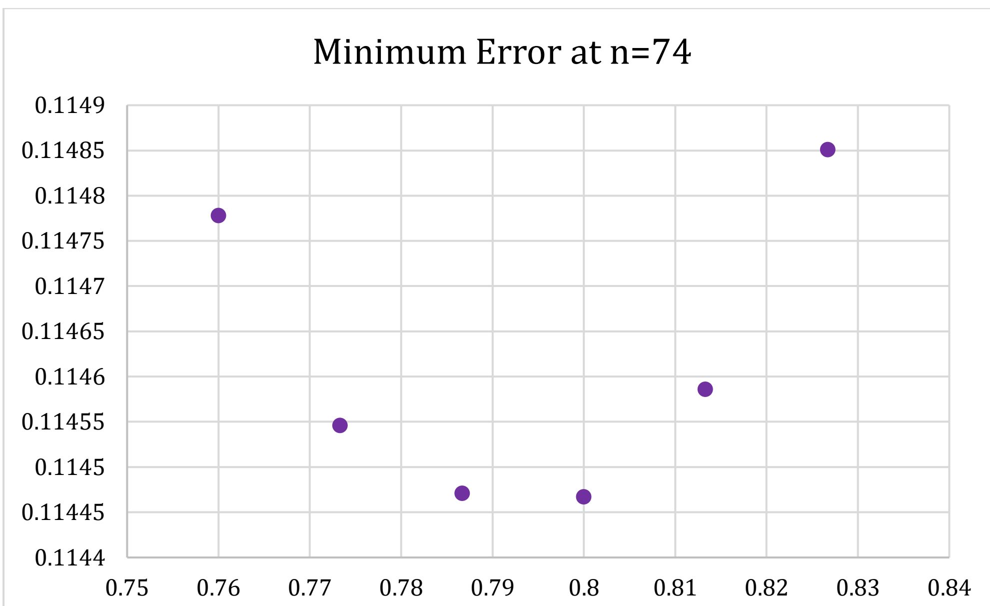


Figure 3.3.9

3.4 Variance and Error Compared

Table 3.4.4 compares which percentiles have minimum variance and minimum error for $n \leq 30$:

Table 3.4.4

N	Error	Variance
1	0.5	0.5
2	0.667	0.667
3	0.75	0.75
4	0.8	0.8
5	0.833	0.833
6	0.714	0.714
7	0.75	0.75
8	0.778	0.778
9	0.8	0.8
10	0.818	0.727
11	0.75	0.75
12	0.769	0.769
13	0.786	0.786
14	0.8	0.8
15	0.813	0.75
16	0.765	0.765
17	0.777	0.777
18	0.789	0.789
19	0.8	0.8
20	0.81	0.762
21	0.773	0.773
22	0.783	0.783
23	0.792	0.792
24	0.8	0.8
25	0.769	0.769
26	0.778	0.778
27	0.785	0.786
28	0.793	0.793
29	0.8	0.8
30	0.774	0.774

The minimum variance and error occurred for the same percentiles for $n \leq 30$ for all but three samples ($n = \{10,15,20\}$). With the exception of expected value of error on these three samples sizes and $n = 5$ the minimum variance and error was at or

immediately below the 80th percentile. An explanation for the three discrepancies can be seen when equations for variance and error can be compared.

For any population variance is:

$$V(x) = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N}(\mu - x_i)^2}{N}$$

Since for unbiased estimators $\theta = \mu$:

$$V(\hat{x}) = \frac{\sum_{i=1}^{N}(\theta - x_i)^2}{N}$$

The expected value of an error for any population can be written as:

$$E(Error) = \frac{\sum_{i=1}^{N}|\theta - x_i|}{N}$$

The relationship between variance and expected value of an error can be better understood with the expected value of an error squared and the expected value of an error quantity squared equations:

$$E((Error)^2) = \frac{\sum_{i=1}^{N}|(\theta - x_i)^2|}{N}$$

$$\left(E(Error)\right)^2 = \frac{(\sum_{i=1}^{n}|\theta - x_i|)^2}{n^2}$$

This means that:

$$V(\hat{x}) = E((Error)^2) \neq \left(E(Error)\right)^2$$

So,

$$\sqrt{(V(\hat{x})} = Standard\ Deviation \neq E(Error)$$

There is no direct comparison between $V(\hat{x})$ and $E(Error)$. Let a, b be percentiles in a sample. As shown in table 3.4.1, sometimes $V_a(x) < V_b(x)$ but $E_b(error) < E_a(error)$. Since both variance and expected value of an error are divided by n, this discrepancy will happen when:

$$E_{a_1}^2 + E_{a_2}^2 + \cdots + E_{a_n}^2 < E_{b_1}^2 + E_{b_2}^2 + \cdots + E_{b_n}^2$$

and

$$E_{a_1} + E_{a_2} + \cdots + E_{a_n} > E_{b_1} + E_{b_2} + \cdots + E_{b_n}$$

where E_{p_k} is an error. At first this might seem paradoxical as $\forall c, d \geq 0 \mid c, d \in \mathbb{R}$, if $c > d$ then $c^2 > d^2$. However, if $0 < c < 1$, then $c > c^2$, so it is possible for $c^2 > d$ or $d > c^2$. So,

when comparing the sum of two sets of positive real numbers, if at least one number in either set is less than 1 it is possible for the sum of the squares of the set with the higher sum to be lower then the sum of the squares of the set with the lower sum. Since the errors are often less then 1, particularly in percentiles that have the lowest expected error, the inconsistencies in the sample of size 10, 15, and 20 is not surprising or problematic. A complete list of variances and expected value of error can be found in appendix B and C.

4. Analyzing Results

4.1 The Benefit of Percentile Matching

Since the expected value of an error deals directly with errors, while variance deals with errors squared, it makes sense to assert that the optimal percentile to use is the one with minimum expected error rather than a minimum variance if there is a discrepancy between the two. This is the benefit to using percentile matching instead of the MVUE $\bar{x}$. There is not an equation for the expected value of an error for $\bar{x}$ for the exponential distribution. With percentile matching, one can know how much error to expect as a percent of the parameter and one cannot do that for $\bar{x}$. While we know that the variance of $\bar{x}$ will always be less that any point estimate created by a percentile, we do not know how the expected value of error for $\bar{x}$ and a percentile point estimate compare. As has already been shown, lower variance does not necessarily equate to lower expected error, so there is no evidence to support that either $\bar{x}$ or an optimal percentile point estimate would always have lower expected error.

4.2 Testing the Results.

In order to test theoretical values, and to see how expected value of error of $\bar{x}$ competed with different percentile point estimates, we simulated 100 samples from an exponential distribution with $\theta = 1$ and $n = 9$. The averages from all 100 samples are shown in the following tables (See appendix D for graphs):

Expected Value of a Percentile

Percentile	Simulated Value	Theoretical Value	Difference
0.1	0.1262	0.111111	0.015089
0.2	0.2751	0.236111	0.038989
0.3	0.4297	0.378968	0.050732
0.4	0.6031	0.545635	0.057465
0.5	0.7916	0.745635	0.045965
0.6	1.0326	0.995635	0.036965
0.7	1.3593	1.32897	0.03033
0.8	1.8393	1.82897	0.01033
0.9	2.7388	2.82897	-0.09017
Mean	1.0218	1	0.0218

Value of an Unbiased Point Estimate

Percenitle	Simulated Value	Theoretical Value	Difference
0.1	1.1361	1	0.1361
0.2	1.1652	1	0.1652
0.3	1.1339	1	0.1339
0.4	1.1053	1	0.1053
0.5	1.0616	1	0.0616
0.6	1.0372	1	0.0372
0.7	1.0228	1	0.0228
0.8	1.0057	1	0.0057
0.9	0.9681	1	-0.0319
Mean	1.0218	1	0.0218

Variance of a Point Estimate

Percentile	Simulated Value	Theoretical Value	Difference
0.1	1.4029	1	0.4029
0.2	0.829	0.502	0.327
0.3	0.4059	0.337	0.0689
0.4	0.3129	0.256	0.0569
0.5	0.2072	0.209	-0.0018
0.6	0.1363	0.18	-0.0437
0.7	0.1441	0.164	-0.0199
0.8	0.1726	0.161	0.0116
0.9	0.1822	0.192	-0.0098
Mean	0.1246	0.111111	0.013489

Expected Value of Error for a Point Estimate

Percentile	Simulated Value	Theoretical Value	Difference
0.1	0.8681	0.735759	0.132341
0.2	0.6794	0.541966	0.137434
0.3	0.4789	0.449857	0.029043
0.4	0.4247	0.394333	0.030367
0.5	0.3591	0.357359	0.001741
0.6	0.293	0.332136	-0.039136
0.7	0.3087	0.316426	-0.007726
0.8	0.3263	0.312149	0.014151
0.9	0.3466	0.335285	0.011315
Mean	0.28		

The expected values for the percentiles are almost exactly what their theoretical values should be. The average difference between the simulated values and the theoretical values is .0217 which is almost exactly the difference between the averages of the means and 1. The 80th percentile produced the most unbiased estimator in the simulated data as its average adjusted point estimate was the closest to 1; it was even closer to 1 than the average of the means. In general, lower percentiles had a higher difference in their simulated variance and error compared to their theoretical values. This happens because lower percentiles have a higher variance in their errors. The 50th, 70th 80th and 90th percentiles had simulated values almost exactly at their theoretical values. The 60th percentile actually had the lowest simulated variance and expected value of error by a small margin and was a bit below its expected value. This can be seen in appendix D. While 100 samples is a large amount, the sample size of 9 is quite small, and so seeing slight deviations in simulated values versus theoretical values is not surprising.

Notice there is no theoretical value for the mean's expected error. That value is not known for the exponential distribution. While the simulated value for the expected error was lower than any of the percentiles, it was only .013 lower than the lowest percentile's simulated values, and .031 lower than the lowest percentile's theoretical value.

4.3 The 80th Percentile

Clearly, there is evidence that percentiles very close to the 80th will produce the best point estimates, but why does this happen at a seemingly random percentile?

When point estimation is done with percentile matching, essentially, a piece of data is assumed to be exactly where it is expected to be in the population's distribution relative to what percentile it is in the sample. The adjustment to make it an unbiased estimator accounts for how large the sample is. We then use that assumption to project what the parameter is. Since lower percentiles have lower variance in how they are distributed, one

might expect the lowest percentile possible to be the best point estimator as it would be consistently closer to what it is expected to be. However, that is clearly not the case. The reason for this is that the higher percentiles deal with variability better when creating a point estimate. Recall that the slope of the error's lines created by percentiles was lower for higher percentiles (Table 3.3.2, Figure 3.3.3). The equation for the slope of an error line is approaching $\frac{1}{-\ln(1-p)}$. Consider the following graph of $\frac{1}{-\ln(1-p)}$ where p is the x-axis:

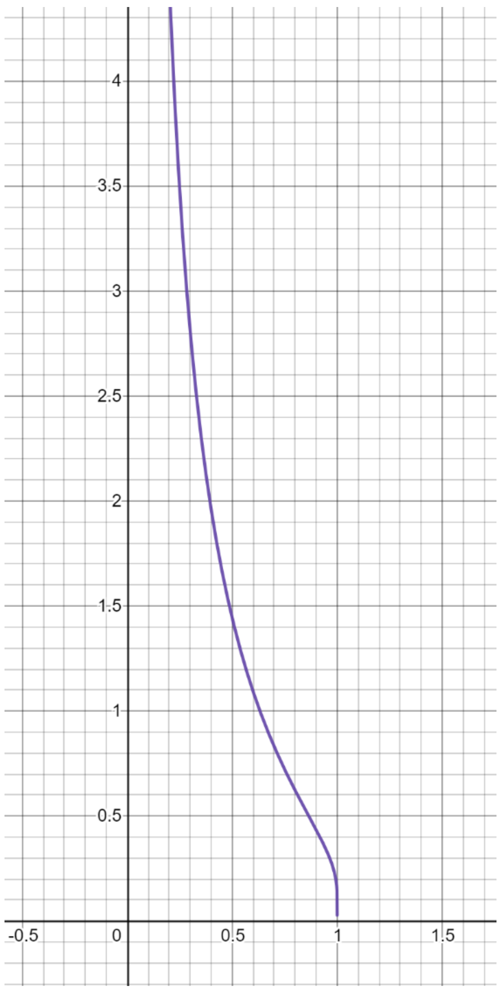


Figure 4.3.1

The purple line represents the slope of an error line for the respective percentile shown on the x-axis. These lower slopes for an error line mean that higher percentiles can have lower error even though they have more variability.

For example, for some exponential sample of size 9 with $\theta = 1$, let its 10^{th} percentile and the 80^{th} percentile both one standard deviation above their expected value. The standard deviation for the 10^{th} percentile is .111113 and the standard deviation for the 80^{th}

percentile is .734689. Clearly the standard deviation for the 80th percentile is significantly higher as the higher percentile has more variability. The resulting point estimates created by the 10th and 80th percentiles would then be:

$$\hat{\theta}_{.1} = \frac{x}{-\ln(1-p)} = \frac{E_{9,1}(x) + 1\sqrt{V_{9,1}(x)}}{-\ln(1-.1)} = \frac{.111111 + .111113}{.105361} = 2.05459$$

$$\hat{\theta}_{.8} = \frac{x}{-\ln(1-p)} = \frac{E_{9,8}(x) + 1\sqrt{V_{9,8}(x)}}{-\ln(1-.8)} = \frac{1.82897 + .734689}{1.60944} = 1.45649$$

While both are higher than 1 (the parameters actual value), as both were a standard deviation above there unbiased expected value, the 80th percentile is a better estimator as it is closer to 1.

A percentile matched point estimate will have a numerator of x and a denominator of $-\ln(1-p)$. Since the expected value of x will be $-\ln(1-p)$ we can expect the numerator and the denominator in the equation of a point estimate to be close together. However, when the values of the numerator and the denominator are higher (i.e. higher percentiles) the ratio between the two is more likely to be closer to 1 even if the difference between the two (variability) is higher. This is why higher percentiles can create point estimates with less error despite having more variability. While lower percentiles have lower variability, higher percentiles deal with this variability more efficiently, and the balance between these two factors appears to be at or very near the 80th percentile.

One final way of understating why this balance occurs near the 80th percentile is to multiply the slope of an error line by the standard deviation of a percentile. While the expected value is the integral of an error line times the PDF of the same percentile, another more intuitive way of producing a similar result is to just take the product of these two factors.

$$Standard\ deviation\ times\ slope\ of\ error\ line = \sqrt{V_{n,k}(x)} \cdot \frac{1}{E_{n,k}(x)} = \frac{\sqrt{V_{n,k}(x)}}{E_{n,k}(x)}$$

Lower values of this product will translate to lower expected error. This is because the value indicates how well a percentile can deal with its variance given how small its error line is. If we do this for a sample of size 9 with $\theta = 1$, the following graph shows the output:

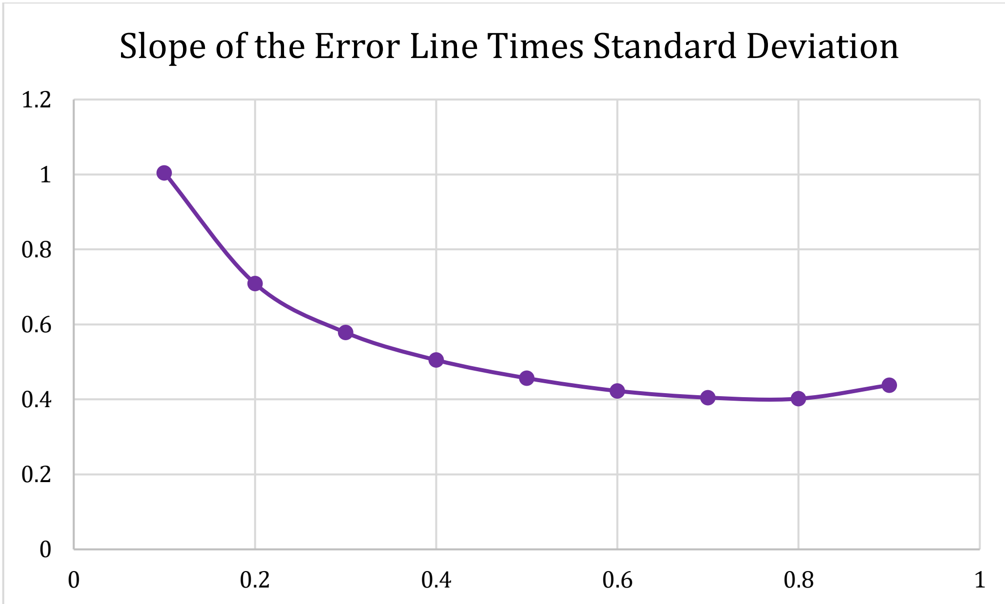


Figure 4.3.2

We can compare this to the output of the theoretical expected value of error (shown in orange).

Figure 4.3.3

While the expected value of error is a more exact measurement of which percentiles are optimal, standard deviation times the slope does give the same ordering for which percentiles work best and shows how the balance of percentile variability and error dictate how good point estimates are for that percentile.

4.4 Percentile Matching in other Distributions

While all of the results in this project focused on the exponential distribution, percentile matching can be used for other distributions as well. Like other methods of point estimation, there are limitations as to when percentile matching can be used. Percentile matching requires a distribution to have a known equation for the CDF so that sample percentiles can be matched to theoretical percentiles, which come from the CDF. This rules out distributions such as the normal distribution which does not have an equation for a CDF. However, sometimes percentile matching is needed when the MVUE is not known or when other methods such as the method of moments cannot be used because a distribution (such as the Cauchy Distribution) does not have moments that can be matched.

While we had successful results in optimizing percentile matching for the exponential distribution, we found it difficult to replicate these results in other distributions. First, it is significantly easier to do percentile matching with distributions that only have one parameter. In order to do it for a two-parameter distribution, two percentiles must be matched, and then the system of two equations can be solved. As already shown, matching more than one percentile will in general lead to increased error. So, when we looked into evaluating percentile matching for other distributions that had more than one parameter, we just set one of the parameters to a constant and evaluated the special case of the distribution. This was done for the Beta, the Cauchy, and the uniform distributions.

Both the Beta and Cauchy distributions had unsuccessful results. To analyze variance or expected value of error of a point estimate created by a percentile, one needs to know a generic formula for the PDF of any order statistic and be able to use that and integration to find the expected value and variance of percentiles. The issues with the Cauchy distribution were with finding a formula for the PDF of all order statistics. While we were able to find this for the Beta distribution, there was trouble integrating to find the expected value and variance of a percentile. This is because the equation for a point estimate for β, when $\alpha = 1$ is $\hat{\beta} = \frac{\ln(p)}{\ln(x)}$. Integrating $\frac{1}{\ln(x)}$ results in the logarithmic integral function. Since this function is an infinite series, it made evaluating it and thus finding the expected value and variance of percentiles in the Beta distribution outside of our scope of research. Another issue with the beta distribution is that the results would not be consistent for different values of β. For the exponential distribution, in every important equation θ was able to be factored out and we could represent the value of expected value of a percentile or expected value of error as a coefficient of θ, and the variance of a percentile and the variance of a point estimate as a coefficient of θ^2. This means that no matter what θ is, the results of which percentiles are optimal are the same, and the results of all percentiles are in the same ratio to one another. It's important

because while we do not know what θ is, we can still obtain meaningful results. This is not able to happen with the parameter in the Beta distribution or Cauchy distribution.

However, the basic concept from the results of the exponential distribution can still be applied to other distributions even if we cannot get precise results. The optimal percentile for a distribution will occur where the variance in a percentile and how well that percentile deals with variance are balanced. For the Beta with $\alpha = 1$, lower percentiles deal with variability better due to the equation $\hat{\beta} = \frac{\ln(p)}{\ln(x)}$. While variability of a percentile will vary depending on the parameter (lower percentiles will have lower variability if $\beta < 1$, and higher percentiles with have lower variability if $\beta > 1$) we can hypothesize that lower percentiles will produce the best point estimates, particularly if we have reason to think $\beta < 1$. This idea can be applied to any distribution on which we need to use percentile matching.

There were productive and unsurprising results for the uniform distribution on $[0, b]$. The exact same steps that were taken to analyze the exponential were applied to the uniform. Consistently the highest percentiles were the optimal ones. This is not surprising because the parameter was given as the upper bound of the distribution, so naturally the percentile that was closest to it produced the best estimate. The uniform distribution was also a case where the parameter was able to be removed from the equation, and results would be the same regardless of the parameter's value.

4.5 Conclusion

The exponential distribution did provide a unique case though, where near the 80$^{\text{th}}$ percentile was the optimal percentile for all sample sizes and for all values of the parameter. This was shown in both the evaluation of the minimum variance and minimum expected error of a point estimate. Finding expected value of error for a point estimate not only indicated which percentile was optimal for any sample size but also produced a number which is not directly represented by either variance or standard deviation, is unknown for the MVUE, but is a helpful one in understanding how good a point estimate is.

Appendix A

This table shows the expected values of a point estimate as a coefficient for θ for a given n, k. Highlighted values are the lowest in their column.

K\N	1	2	3	4	5	6	7	8	9	10
1	1.4427	1.23315	1.15869	1.12036	1.09696	1.08119	1.06984	1.06127	1.05458	1.04921
2		1.36536	1.20225	1.14194	1.10984	1.08974	1.07592	1.06582	1.05811	1.05203
3			1.32247	1.1823	1.13011	1.10195	1.08408	1.07167	1.0625	1.05545
4				1.29445	1.16814	1.12121	1.09576	1.07951	1.06814	1.0597
5					1.27435	1.15744	1.11422	1.09075	1.07572	1.06517
6						1.25905	1.149	1.10854	1.08659	1.07252
7							1.2469	1.14213	1.10382	1.08307
8								1.23695	1.1364	1.09981
9									1.22861	1.13153
10										1.22147

K\N	11	12	13	14	15	16	17	18	19	20
1	1.0448	1.04111	1.03799	1.0353	1.03297	1.031	1.029	1.028	1.026	1.025
2	1.0471	1.04303	1.03961	1.03669	1.03418	1.032	1.03	1.028	1.027	1.025
3	1.04984	1.04527	1.04148	1.03828	1.03554	1.033	1.031	1.029	1.028	1.026
4	1.05316	1.04794	1.04367	1.04011	1.0379	1.035	1.032	1.03	1.029	1.027
5	1.05729	1.05117	1.04627	1.04225	1.03888	1.036	1.034	1.031	1.03	1.028
6	1.06261	1.0552	1.04943	1.04479	1.04097	1.038	1.035	1.033	1.031	1.029
7	1.06976	1.06038	1.05336	1.04788	1.04346	1.04	1.037	1.034	1.032	1.03
8	1.08004	1.06736	1.05842	1.05173	1.04649	1.042	1.039	1.036	1.033	1.031
9	1.09636	1.0774	1.06525	1.05668	1.05026	1.045	1.041	1.038	1.035	1.033
10	1.12732	1.09334	1.07506	1.06337	1.05512	1.049	1.044	1.04	1.037	1.034
11	1.21529	1.12363	1.09068	1.07299	1.06168	1.054	1.048	1.043	1.039	1.036
12		1.20985	1.12037	1.08831	1.07113	1.06	1.052	1.047	1.042	1.038
13			1.20503	1.11745	1.08617	1.069	1.059	1.051	1.046	1.041
14				1.2007	1.11483	1.084	1.068	1.058	1.05	1.045
15					1.1968	1.112	1.082	1.067	1.056	1.049
16						1.193	1.11	1.081	1.065	1.055
17							1.19	1.102	1.079	1.064
18								1.187	1.106	1.078
19									1.184	1.105
20										1.182

Appendix B

This table shows the variance of an unbiased point estimator as a coefficient of θ^2 . Each value is an output from Equation 3.2.1 when the respective n (top row) and k (first column) are the inputs. The highlighted values are the minimal variance for that sample size. Their respective n and k values are columns 1 and 2 in Table 3.24.

K\N	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2		0.5560	0.5200	0.5100	0.5060	0.5040	0.5030	0.5020	0.5020	0.5010
3			0.4050	0.3610	0.3480	0.3430	0.3400	0.3380	0.3370	0.3360
4				0.3280	0.2814	0.2670	0.2610	0.2580	0.2560	0.2540
5					0.2810	0.2340	0.2190	0.2130	0.2090	0.2070
6						0.2480	0.2020	0.1870	0.1800	0.1760
7							0.2250	0.1790	0.1640	0.1570
8								0.2070	0.1610	0.1470
9									0.1920	0.1480
10										0.1810

K\N	11	12	13	14	15	16	17	18	19	20
1	1	1	1	1	1	1	1	1	1	1
2	0.5010	0.5010	0.5008	0.5006	0.5005	0.5005	0.5005	0.5004	0.5004	0.5003
3	0.3360	0.3350	0.3349	0.3346	0.3344	0.3343	0.3342	0.3341	0.3340	0.3339
4	0.2540	0.2530	0.2520	0.2520	0.2517	0.2515	0.2513	0.2512	0.2510	0.2509
5	0.2050	0.2040	0.2030	0.2030	0.2020	0.2020	0.2018	0.2016	0.2014	0.2012
6	0.1740	0.1720	0.1710	0.1710	0.1699	0.1694	0.1691	0.1687	0.1680	0.1683
7	0.1530	0.1510	0.1490	0.1480	0.1470	0.1460	0.1459	0.1455	0.1450	0.1449
8	0.1400	0.1360	0.1330	0.1320	0.1300	0.1296	0.1289	0.1283	0.1280	0.1275
9	0.1330	0.1260	0.1220	0.1200	0.1180	0.1170	0.1159	0.1152	0.1146	0.1142
10	0.1370	0.1230	0.1160	0.1120	0.1090	0.1070	0.1060	0.1051	0.1040	0.1037
11	0.1710	0.1280	0.1140	0.1070	0.1030	0.1003	0.0985	0.0972	0.0962	0.0954
12		0.1630	0.1200	0.1060	0.0990	0.0955	0.0929	0.0911	0.0898	0.0888
13			0.1550	0.1140	0.1000	0.0932	0.0893	0.0867	0.0849	0.0835
14				0.1490	0.1080	0.0945	0.0879	0.0839	0.0813	0.0795
15					0.1440	0.1030	0.0898	0.0832	0.0793	0.0767
16						0.1390	0.9877	0.0856	0.0791	0.0750
17							0.1342	0.0949	0.0820	0.0755
18								0.1302	0.0915	0.0787
19									0.1266	0.0883
20										0.1233

K\N	21	22	23	24	25	26	27	28	29	30
1	1	1	1	1	1	1	1	1	1	1
2	0.5003	0.5003	0.5002	0.5002	0.5002	0.5002	0.5002	0.5002	0.5002	0.5001
3	0.3339	0.3338	0.3338	0.3338	0.3337	0.3337	0.3337	0.3336	0.3336	0.3336
4	0.2508	0.2507	0.2507	0.2506	0.2506	0.2505	0.2505	0.2504	0.2504	0.2503
5	0.2011	0.2010	0.2009	0.2008	0.2008	0.2007	0.2006	0.2006	0.2005	0.2005
6	0.1681	0.1680	0.1678	0.1677	0.1676	0.1676	0.1675	0.1674	0.1674	0.1673
7	0.1447	0.1445	0.1443	0.1442	0.1441	0.1440	0.1439	0.1438	0.1437	0.1437
8	0.1270	0.1270	0.1268	0.1266	0.1265	0.1263	0.1262	0.1261	0.1260	0.1260
9	0.1138	0.1135	0.1133	0.1130	0.1129	0.1127	0.1126	0.1124	0.1123	0.1122
10	0.1033	0.1029	0.1370	0.1023	0.1021	0.1019	0.1017	0.1015	0.1014	0.1013
11	0.0948	0.0943	0.0939	0.0936	0.0933	0.0931	0.0929	0.0927	0.0926	0.0924
12	0.0880	0.0874	0.0869	0.0865	0.0862	0.0859	0.0856	0.0854	0.0852	0.0851
13	0.0825	0.0820	0.0811	0.0806	0.0802	0.0796	0.0796	0.0793	0.0791	0.0789
14	0.0782	0.0772	0.0764	0.0758	0.0753	0.0748	0.0745	0.0742	0.0739	0.0737
15	0.0749	0.0736	0.0726	0.0718	0.0711	0.0706	0.0717	0.0698	0.0695	0.0692
16	0.0727	0.0709	0.0695	0.0685	0.0677	0.0671	0.0665	0.0661	0.0657	0.0654
17	0.0717	0.0691	0.0673	0.0660	0.0649	0.0641	0.0635	0.0629	0.0625	0.0621
18	0.0723	0.0685	0.0659	0.0641	0.0627	0.0618	0.0610	0.0603	0.0598	0.0593
19	0.0757	0.0694	0.0656	0.0631	0.0613	0.0599	0.0589	0.0581	0.0575	0.0569
20	0.0855	0.0730	0.0668	0.0630	0.0605	0.0587	0.0574	0.0564	0.0556	0.0549
21	0.1203	0.0829	0.0706	0.0644	0.0607	0.0582	0.0564	0.0551	0.0541	0.0532
22		0.1175	0.0806	0.0684	0.0622	0.0585	0.0561	0.0543	0.0530	0.0519
23			0.1149	0.0784	0.0663	0.0602	0.0566	0.0541	0.0524	0.0510
24				0.1125	0.0764	0.0644	0.0584	0.0548	0.0523	0.0506
25					0.1103	0.0745	0.0627	0.0567	0.0531	0.0507
26						0.1082	0.0728	0.0611	0.0552	0.0516
27							0.1062	0.0712	0.0596	0.0537
28								0.1044	0.0697	0.0582
29									0.1026	0.0682
30										0.1010

Appendix C

This table is the values for the expected error of an unbiased point estimate as a coefficient of θ. Each value is an output from Equation 3.3.2 when the respective n (top row) and k (first column) are the inputs. The highlighted values are the minimal error for that sample size. Their respective n and k values are columns 1 and 2 in Table 3.3.3.

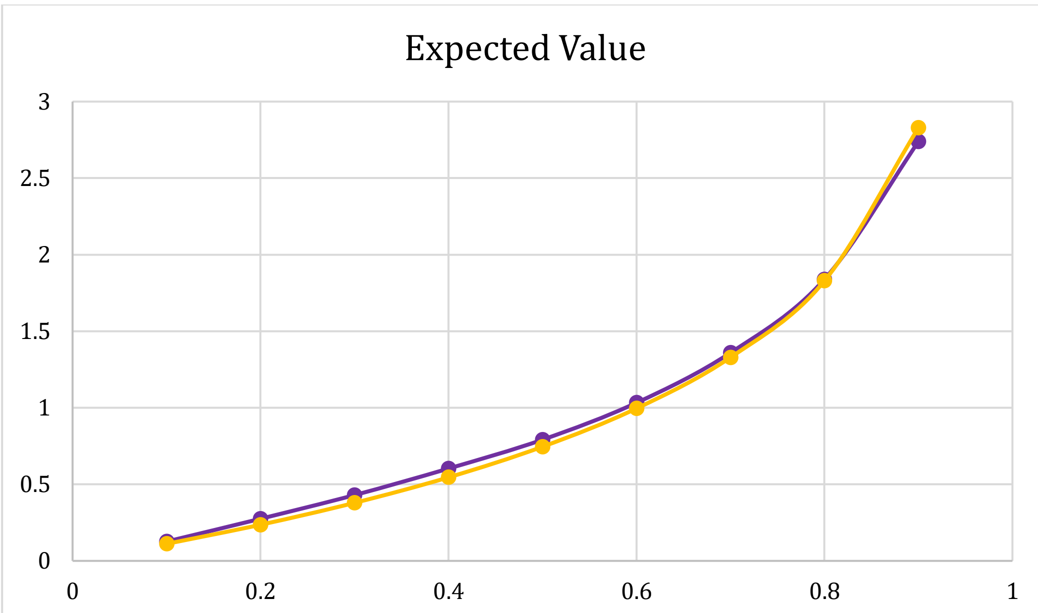
K\N	1	2	3	4	5	6	7	8	9	10
1	**0.736**	0.736	0.736	0.736	0.736	0.736	0.736	0.736	0.736	0.736
2		**0.562**	0.549	0.545	0.544	0.543	0.542	0.542	0.542	0.542
3			**0.483**	0.462	0.455	0.453	0.451	0.450	0.450	0.449
4				**0.436**	0.410	0.401	0.398	0.396	0.394	0.394
5					**0.404**	**0.374**	0.365	0.360	0.357	0.356
6						0.380	**0.348**	0.337	0.332	0.329
7							0.362	**0.328**	0.316	0.311
8								0.347	**0.312**	0.300
9									0.335	**0.299**
10										0.325

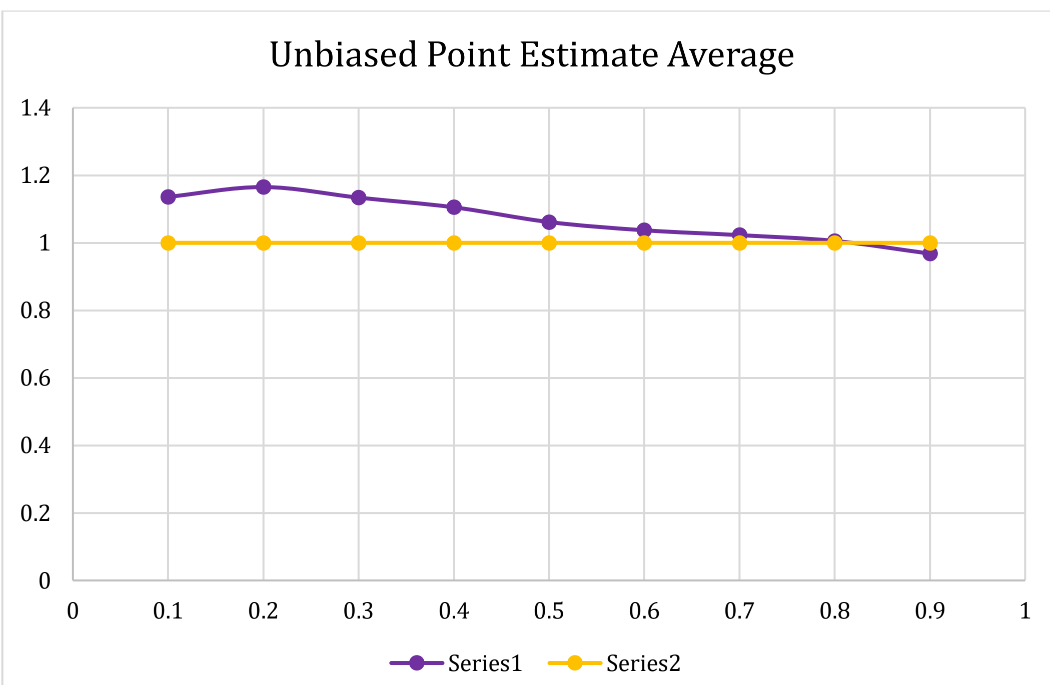
K\N	11	12	13	14	15	16	17	18	19	20
1	0.736	0.736	0.736	0.736	0.736	0.736	0.736	0.736	0.736	0.736
2	0.542	0.542	0.542	0.542	0.542	0.542	0.542	0.541	0.542	0.541
3	0.449	0.449	0.449	0.449	0.449	0.449	0.449	0.448	0.448	0.448
4	0.393	0.393	0.392	0.392	0.392	0.392	0.392	0.391	0.391	0.391
5	0.355	0.354	0.353	0.353	0.353	0.352	0.352	0.352	0.352	0.352
6	0.327	0.326	0.325	0.324	0.324	0.324	0.323	0.323	0.323	0.323
7	0.307	0.305	0.304	0.303	0.302	0.301	0.301	0.300	0.300	0.300
8	0.293	0.290	0.287	0.286	0.285	0.284	0.283	0.282	0.282	0.282
9	**0.285**	0.279	0.275	0.273	0.271	0.270	0.269	0.268	0.267	0.267
10	0.288	**0.274**	0.267	0.263	0.260	0.258	0.257	0.256	0.255	0.254
11	0.316	0.278	**0.264**	0.257	0.253	0.250	0.248	0.246	0.245	0.244
12		0.308	0.270	**0.255**	0.248	0.243	0.240	0.238	0.237	0.236
13			0.302	0.262	**0.248**	**0.240**	0.235	0.232	0.230	0.228
14				0.295	0.256	0.241	**0.233**	0.228	0.225	0.223
15					0.290	0.250	0.235	**0.227**	0.222	0.219
16						0.285	0.245	0.230	**0.221**	0.216
17							0.280	0.240	0.225	**0.216**
18								0.276	0.236	0.220
19									0.272	0.232
20										0.269

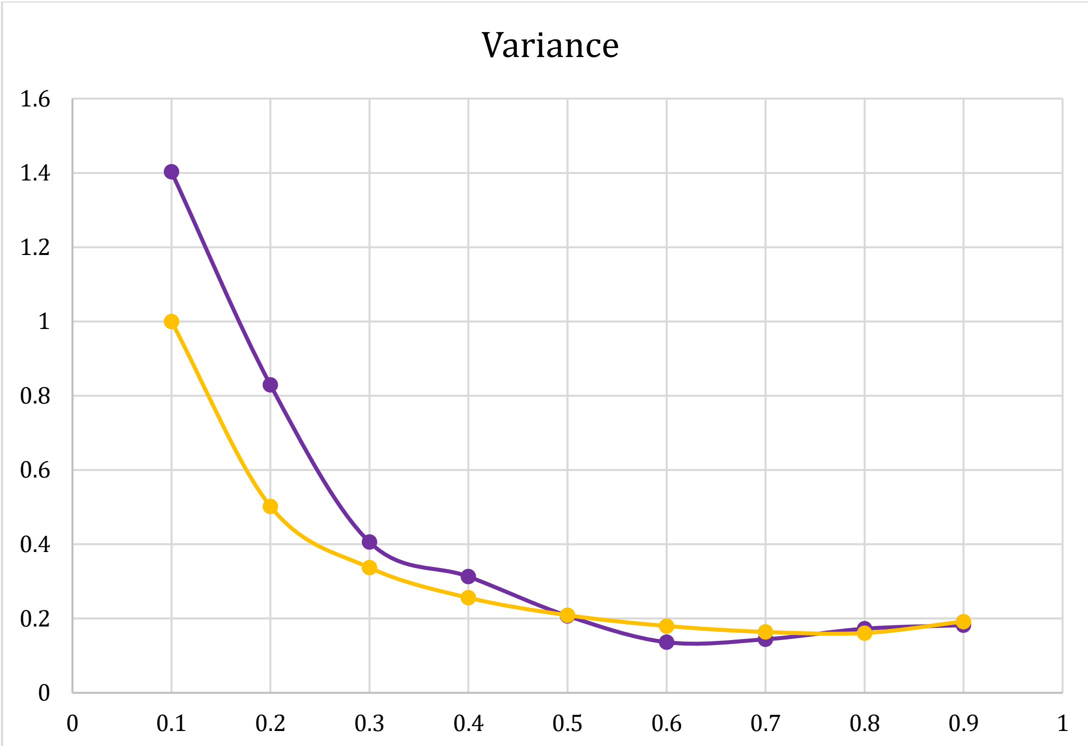
K\N	21	22	23	24	25	26	27	28	29	30
1	0.736	0.736	0.736	0.736	0.736	0.736	0.736	0.736	0.736	0.736
2	0.541	0.541	0.541	0.541	0.541	0.541	0.541	0.541	0.541	0.541
3	0.448	0.448	0.448	0.448	0.448	0.448	0.448	0.448	0.448	0.448
4	0.391	0.391	0.391	0.391	0.391	0.391	0.391	0.391	0.391	0.391
5	0.352	0.352	0.352	0.352	0.351	0.351	0.351	0.351	0.351	0.351
6	0.322	0.322	0.322	0.322	0.322	0.322	0.322	0.322	0.322	0.322
7	0.300	0.299	0.299	0.299	0.299	0.299	0.299	0.299	0.299	0.299
8	0.281	0.281	0.281	0.281	0.281	0.280	0.280	0.280	0.280	0.280
9	0.266	0.266	0.266	0.266	0.265	0.265	0.265	0.265	0.265	0.265
10	0.254	0.253	0.253	0.253	0.253	0.252	0.252	0.252	0.252	0.252
11	0.243	0.243	0.242	0.242	0.242	0.241	0.241	0.241	0.241	0.241
12	0.235	0.234	0.233	0.233	0.232	0.232	0.232	0.231	0.231	0.231
13	0.227	0.226	0.225	0.225	0.224	0.224	0.223	0.223	0.223	0.222
14	0.221	0.220	0.219	0.218	0.217	0.217	0.216	0.216	0.215	0.215
15	0.216	0.215	0.213	0.212	0.211	0.210	0.210	0.209	0.209	0.209
16	0.213	0.210	0.209	0.207	0.206	0.205	0.204	0.204	0.203	0.203
17	0.211	0.208	0.205	0.203	0.202	0.201	0.200	0.199	0.198	0.198
18	0.212	0.206	0.203	0.200	0.198	0.197	0.196	0.195	0.194	0.193
19	0.216	0.207	0.202	0.198	0.196	0.194	0.192	0.191	0.190	0.189
20	0.228	0.212	0.204	0.198	0.194	0.192	0.190	0.188	0.187	0.186
21	0.266	0.225	0.209	0.200	0.194	0.191	0.188	0.186	0.184	0.183
22		0.265	0.221	0.205	0.197	0.191	0.187	0.184	0.182	0.181
23			0.260	0.218	0.202	0.193	0.188	0.184	0.181	0.179
24				0.257	0.216	0.199	0.190	0.185	0.181	0.178
25					0.254	0.213	0.197	0.188	0.182	0.178
26						0.252	0.210	0.194	0.185	0.179
27							0.250	0.208	0.192	0.183
28								0.247	0.206	0.194
29									0.245	0.204
30										0.243

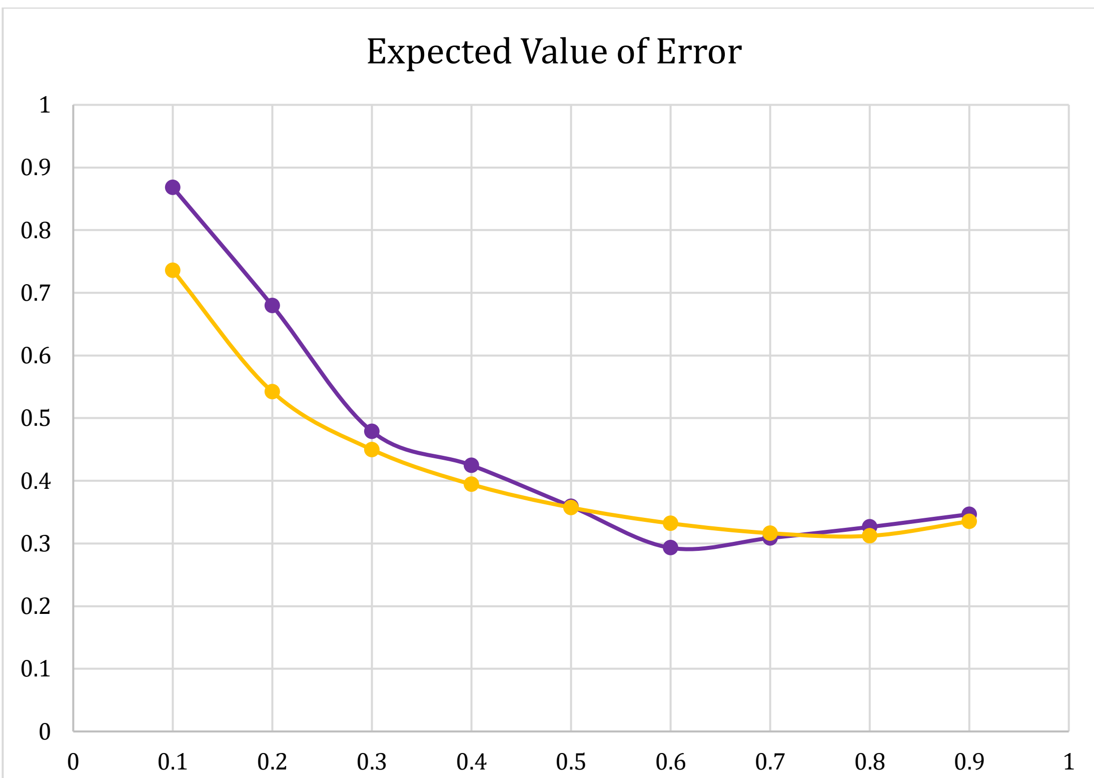
Appendix D

These four graphs compare the simulated averages to the theoretical values for the 9 percentiles from 100 samples of size 9 from an exponential population with $\theta = 1$. For all graphs, the purple line represents the simulated values, and the orange represents the theoretical values.






Variance
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Expected Value of Error
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

References

Jowett, G. H. “The Exponential Distribution and Its Applications.” *The Incorporated Statistician*, vol. 8, no. 2, 1958, pp. 89–95. *JSTOR*, www.jstor.org/stable/2986561. Accessed 23 July 2021.

Sampath, S, and K. Anjana. “Percentile Matching Estimation of Uncertainty Distribution.” *Journal of Uncertainty Analysis and Applications*, vol. 4, no. 1, 2016, doi:10.1186/s40467-016-0049-9.

Schoonjans, Frank, et al. “Estimation of Population Percentiles.” *Epidemiology*, vol. 22, no. 5, 2011, pp. 750–751. *JSTOR*, www.jstor.org/stable/23047978. Accessed 13 July 2021.

Triveri, James D. “Smoothed Empirical Percentile for Percentile Matching.” *The Pleasure of Finding Things Out*, 1 July 2017, www.jtrive.com/smoothed-empirical-percentile-for-percentile-matching.html.

Wackerley, Mendenhall, and Schea§er (2008). Mathematical Statistics with Applications (7th Edition), Thomson: Brooks/Cole.