8-2005

The Interquartile Range: Theory and Estimation.

Dewey Lonzo Whaley
East Tennessee State University

The Interquartile Range: Theory and Estimation

A Thesis

Presented to the Faculty of the Department of Mathematics

East Tennessee State University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Mathematical Sciences

---

by

Dewey L. Whaley III

August, 2005

Robert Price, Ph.D., Chair

Edith Seier, Ph.D.

Robert Gardner, Ph.D

Keywords: Interquartile Range, Probability Distribution, Order Statistics,

Bootstrapping

ABSTRACT

The Interquartile Range: Theory and Estimation

by

Dewey L. Whaley III

The interquartile range (IQR) is used to describe the spread of a distribution. In an introductory statistics course, the IQR might be introduced as simply the "range within which the middle half of the data points lie." In other words, it is the distance between the two quartiles, $IQR = Q_3 - Q_1$. We will compute the population IQR, the expected value, and the variance of the sample IQR for various continuous distributions. In addition, a bootstrap confidence interval for the population IQR will be evaluated.

DEDICATION

I would like to dedicate this thesis to my wife, Jenny, for her encouragement and support during the writing of this thesis. I would also like to thank my parents, Martha and D.L. Whaley, who taught me to finish what I started and especially for pushing me to succeed.

ACKNOWLEDGMENTS

I would like to take this opportunity to thank the chair of my thesis committee, Dr. Robert Price, for his time and support. I also appreciate the time given by the other members of this committee, Dr. Edith Seier and Dr. Robert Gardner.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

One goal in statistical analysis is to describe the center and spread of a distribution with numbers. One of the most common statistics used to describe the center of the distribution is the "mean". The sample mean, $\bar{x}$, is given by the following equation,

$$\bar{x} = \frac{1}{n} \sum x_i, \tag{1}$$

where x_i denotes the ith observed measurement.

The central tendency is a good starting point for describing a sample of data. By itself, however, the central tendency does not provide enough information. For example, assume that two companies want to compare the salaries of their top five management positions. These salaries are given in the following table.

Table 1: Salaries for Two Hypothetical Companies

Title	**Company A**	**Company B**
C.E.O.	80,000	150,000
C.F.O.	75,000	90,000
C.O.O.	75,000	50,000
C.T.O.	75,000	45,000
C.I.O.	70,000	40,000

For this example, the sample mean is equal to \$75,000 for both Companies A and B. It is obvious, however, that these samples are very different. The spread or variability of salaries for Company B is much larger. In this situation, the sample mean alone can sometimes be misleading. It does not tell the whole story. For this reason, a measurement of the "spread" is also a valuable tool in describing a data set.

The purpose for measuring spread is to determine the variability of the sample data. One measure of spread considers the distance of an observation from the center of a distribution. For example, the "deviation" of an observation x from the mean ($\bar{x}$) is $(x - \bar{x})$; the difference between the two values [1]. At first, one might think that the best way to calculate the spread would be to take the average of the deviations. It can be shown, however, that the sum or the average of the deviations for any data set will always be zero, That is, $\sum(x - \bar{x}) = 0$. Because of this, summary measures for the spread typically use the squared deviations or their absolute values. The more common of the two methods is to square the deviations. The average of the squared deviations is called the variance, s^2, and the square root of the variance is then called the standard deviation, s. The mean and standard deviation have good properties and are relatively easy to compute. The mean and standard deviation are also particularly useful descriptive statistics when the distribution is symmetric and no outliers are present.

The mean, however, is not always the best choice for describing the center of a data set. Agresti and Franklin describe the mean as "not resistant" [1]. That is to say that the mean is greatly influenced by outliers or distributions that are skewed. Fortunately, other statistics can be used when describing a data set. Consider the example from Table 1. The salaries for Company A exhibit a symmetric distribution, and the mean of \$75,000 is a good measure of the center of the data. The mean for Company B is also \$75,000, but three of the five numbers are less than \$75,000. In this case, the mean may not be the best description of the center. Company B from the example above is not symmetrical. That is, it has a skewed distribution. In the

case of a skewed distribution, the median may be a better measure of central tendency than the mean. As Agresti might say, the median is "resistant". It is affected very little by outliers.

Let $x_1, x_2, x_3, \ldots, x_n$ denote an observed sample of n measurements. By putting the observations in order from least to greatest, we have, $y_1, y_2, y_3, \ldots, y_n$ such that $y_1 \leq y_2 \leq y_3 \leq \cdots \leq y_n$. Once the data is ordered, it is easy to find the minimum, maximum, and quartiles. To find the second quartile or the median, the data from a random sample, $x_1, x_2, x_3, \ldots, x_n$, must be placed in order, as mentioned above, to yield $y_1, y_2, y_3, \ldots, y_n$. The sample median is the number such that 50% of the observations are smaller and 50% are larger. Formally, the sample median, M, is defined as

$$M = \begin{cases} y_{(n+1)/2}, & \text{if } n \text{ is odd} \\ (y_{n/2} + y_{n/2+1})/2, & \text{if } n \text{ is even} \end{cases} \tag{2}$$

Even though the median ignores the numerical values from nearly all of the data, it can sometimes provide a more accurate measure of the central tendency. The median of the salaries from company B is \$50,000 which seems to be a better measure of the center for the top five salaries of the company. As mentioned earlier in the discussion of the mean, the measure of central tendency does not provide enough helpful information about the sample. In such cases, it is important to examine the "spread" or "dispersion" of the data set.

The standard deviation measures spread about the mean. Therefore, it is not practical to calculate the standard deviation when using the median as the measure of central tendency. Other statistics may be more useful when calculating the spread

about the median. One statistic that is often used to measure the spread is to calculate the range. The range is found by subtracting the smallest value in the sample, y_1, from the largest value, y_n. The problem with the range is that it shares the worst properties of the mean and the median. Like the mean, it is not resistant. Any outlier in any direction will significantly influence the value of the range. Like the median, it ignores the numerical values of most of the data. That is not to say that the range does not provide any useful information and it is a relatively easy statistic to compute. In order to avoid the problem of dealing with the outliers, however, we can calculate a different measure of dispersion called the interquartile range (IQR). The interquartile range can be found by subtracting the first quartile value (q_1) from the third quartile value (q_3). For a sample of observations, we define q_1 to be the order statistic below which 25% of the data lies. Similarly, q_3 is defined to be the order statistic, below which 75% of the data lies. According to Hogg and Tanis,

> If $0 < p < 1$ the $(100p)$th sample has "approximately" np sample observations less than it and also $n(1-p)$ sample observations greater than it. One way of achieving this is to take the $(100p)$th sample percentiles as the $(n+1)p$th order statistic provided that $(n+1)p$ is an integer. If $(n+1)p$ is not an integer but is equal to r plus some proper fraction say $\frac{a}{b}$, use a weighted average of the rth and the $(r+1)$st order statistics.[5]

Based on the Hogg and Tanis method, we must identify the different cases in which $(n+1)p$ is an integer and when it is not an integer. In calculating the interquartile range, we will be working with p equal to .25 and .75 and four different cases for the value of n. The first case is the only one in which $(n+1).25$ and $(n+1).75$ are integers.

If $n = 4m - 1$ and m is an integer, then $(n + 1).25 = ((4m - 1) + 1).25 = m$ and $(n + 1).75 = ((4m - 1) + 1).75 = 3m$ which yield the mth and $3m$th order statistics. Hence, $q_1 = y_m$ and $q_3 = y_{3m}$.

The second case we consider is when $n = 4m$. To find the first quartile, we work with $(n + 1).25 = (4m + 1).25 = m + \frac{1}{4}$. Hogg and Tanis say in this case we must take a weighted average of the mth and $(m + 1)$th order statistics. That is,

$$q_1 = \left(1 - \frac{1}{4}\right) y_m + \left(\frac{1}{4}\right) y_{m+1} \tag{3}$$

or simply,

$$q_1 = \left(\frac{3}{4}\right) y_m + \left(\frac{1}{4}\right) y_{m+1}. \tag{4}$$

For the third quartile, we have $(n + 1).75 = 3m + \frac{3}{4}$ and by Hogg and Tanis' method of weighted averages [5] we find that

$$q_3 = \left(1 - \frac{3}{4}\right) y_{3m} + \left(\frac{3}{4}\right) y_{3m+1} \tag{5}$$

or simply,

$$q_3 = \left(\frac{1}{4}\right) y_{3m} + \left(\frac{3}{4}\right) y_{3m+1}. \tag{6}$$

The third case is when $n = 4m + 2$. Again, we will calculate q_1 and q_3. We have $(n + 1).25 = (4m + 2 + 1).25 = m + \frac{3}{4}$ and first quartile is

$$q_1 = \left(\frac{1}{4}\right) y_m + \left(\frac{3}{4}\right) y_{m+1}. \tag{7}$$

For the third quartile, we have $(n + 1).75 = (4m + 2 + 1).75 = 3m + 2 + \frac{1}{4}$ and since m is an integer, $3m + 2$ is also an integer. In the context of Hogg and Tanis, the rth order statistic will correspond to the $3m + 2$ order statistic. Then by the method of

weighted averages,

$$q_3 = \left(\frac{3}{4}\right) y_{3m+2} + \left(\frac{1}{4}\right) y_{3m+3} \tag{8}$$

Finally, the fourth case is for $n = 4m + 1$. The first and third quartiles are then found using the same method that we used above. In short we have,

$$q_1 = \left(\frac{1}{2}\right) y_m + \left(\frac{1}{2}\right) y_{m+1} \tag{9}$$

and

$$q_3 = \left(\frac{1}{2}\right) y_{3m+1} + \left(\frac{1}{2}\right) y_{3m+2}. \tag{10}$$

For the salary example given in Table 1, we now compute the IQR for each of the companies. For Company A we see that $n = 5$ and this falls into the $n = 4m + 1$ category with $m = 1$. Thus we have,

$$\begin{aligned} q_1 &= \left(\frac{1}{2}\right) y_1 + \left(\frac{1}{2}\right) y_2, \\ q_3 &= \left(\frac{1}{2}\right) y_4 + \left(\frac{1}{2}\right) y_5. \end{aligned} \tag{11}$$

So, for Company A in the example,

$$\begin{aligned} q_1 &= \left(\frac{1}{2}\right) (\$70,000) + \left(\frac{1}{2}\right) (\$75,000) \\ &= \$72,500 \\ q_3 &= \left(\frac{1}{2}\right) (\$75,000) + \left(\frac{1}{2}\right) (\$80,000) \\ &= \$77,500. \end{aligned} \tag{12}$$

Finally, to calculate the IQR, we simply take the difference,

$$\begin{aligned} IQR &= \$77,500 - \$72,500 \\ &= \$5,000. \end{aligned} \tag{13}$$

This means that for the middle half of the salaries in Company A, \$5,000 is the distance between the largest and smallest salaries. Similarly, we find the IQR for Company B to be \$77,500. Obviously Company B's salaries are much more "spread out" than Company A's.

Theoretical properties of the IQR will be explored in Chapters 2, 3, and 4. In Chapter 2 we compute the population IQR for six continuous distributions: Uniform, Exponential, Normal, Gamma, Lognormal, and Weibull. Chapters 3 and 4 compute the expected value and variance of the sample IQR for three continuous distributions: Uniform, Exponential, and Chi-Square.

In the final chapter, estimation of the expected value and variance of the sample IQR using bootstrapping is investigated. In addition, a bootstrap confidence interval for the population IQR called the percentile method will be given. A simulation study will check the accuracy of the percentile method.

2 THE POPULATION INTERQUARTILE RANGE

The population IQR for a continuous distribution is defined to be

$$IQR = Q_3 - Q_1 \tag{14}$$

where Q_3 and Q_1 are found by solving the following integrals

$$.75 = \int_{-\infty}^{Q_3} f(x)\,dx \text{ and } .25 = \int_{-\infty}^{Q_1} f(x)\,dx. \tag{15}$$

The function $f(x)$ is continuous over the support of X that satisfies the two properties,

$$(i) : f(x) \geq 0 \text{ and } (ii) : \int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{16}$$

In this section, we will find the population IQR for the following distributions: uniform, exponential, normal, gamma, lognormal, and Weibull.

2.1 The Uniform Distribution

An important distribution in generating pseudo-random numbers is the uniform distribution. A random variable X is said to have a continuous uniform probability distribution on the interval (θ_1, θ_2) if and only if the probability density function (p.d.f.) of X is

$$f(x) = \frac{1}{\theta_2 - \theta_1}, \quad \theta_1 \leq x \leq \theta_2. \tag{17}$$

This distribution is sometimes referred to as rectangular. Figure 1 is an illustration of a uniform density function when $\theta_1 = 0$ and $\theta_2 = 1$.

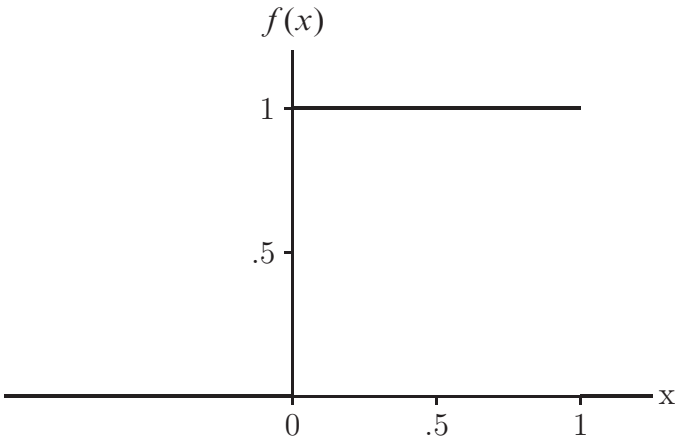

Figure 1: Uniform ($\theta_1 = 0, \theta_2 = 1$) p.d.f.

Applying equations (14), (15), and (17) we find the following:

$$.75 = \int_{\theta_1}^{Q_3} \frac{1}{\theta_2 - \theta_1} dx. \tag{18}$$

So, after integrating and solving for Q_3 we find

$$Q_3 = .75(\theta_2 - \theta_1) + \theta_1. \tag{19}$$

Similarly,

$$Q_1 = .25(\theta_2 - \theta_1) + \theta_1. \tag{20}$$

Now that we have the parameters Q_1 and Q_3, it is easy to take the difference which yields the population IQR,

$$IQR_{\text{unif}} = .5(\theta_2 - \theta_1). \tag{21}$$

2.2 The Exponential Distribution

The exponential distribution is a probability model for waiting times. For example, consider a process in which we count the number of occurrences in a given interval

and this number is a realization of a random variable with an approximate Poisson distribution. Because the number of occurrences is a random variable it follows that the waiting times between occurrences is also a random variable. If λ is the mean number of occurrences in a unit interval, then $\beta = 1/\lambda$ is the mean time between events.[5] The density function for the exponential distribution is given by

$$f(x) = \frac{1}{\beta} e^{-x/\beta},\ x > 0. \tag{22}$$

The shape of an exponential distribution is skewed right. Figure 2 is an illustration of an exponential density function when $\beta = 1$.

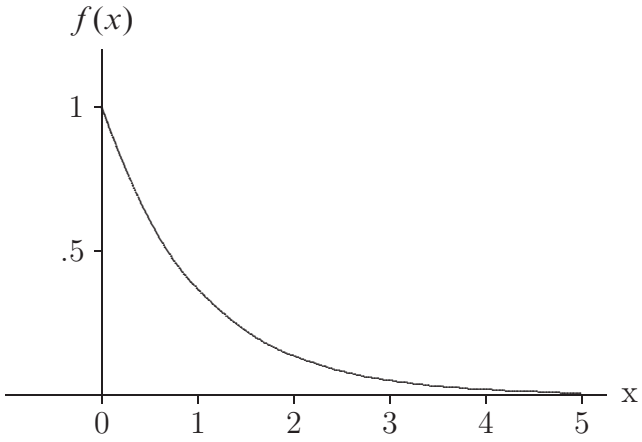

Figure 2: Exponential ($\beta = 1$) p.d.f.

This time we integrate from 0 to Q_3,

$$.75 = \int_0^{Q_3} \frac{1}{\beta} e^{-x/\beta} dx. \tag{23}$$

After solving for Q_3 the above equation becomes

$$Q_3 = -\beta \ln .25, \tag{24}$$

Q_1 can then be found in a similar fashion,

$$Q_1 = -\beta \ln .75. \tag{25}$$

Finally, by subtracting Q_1 from Q_3, we see that the IQR parameter is

$$IQR_{\text{exp}} = 1.0986\beta. \tag{26}$$

2.3 The Normal Distribution

An important distribution in statistical data analysis is the normal distribution. Many random variables (e.g. heights by gender, birth weights of babies in the U.S., blood pressures in an age group) can be approximated by the normal distribution. The Central Limit Theorem motivates the use of the normal distribution. The normal distribution is given by the p.d.f.,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty. \tag{27}$$

The shape of a normal distribution is symmetric about the mean μ. Figure 3 is an illustration of a standard normal density function ($\mu = 0, \sigma = 1$).

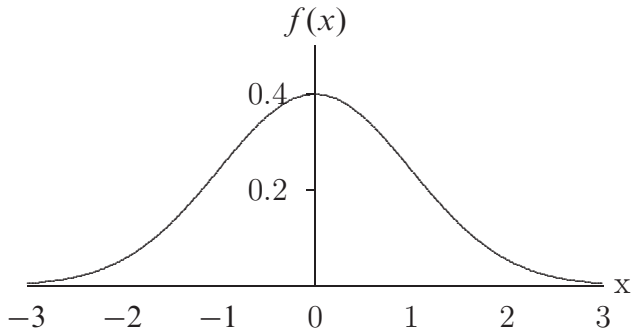

Figure 3: Normal ($\mu = 0, \sigma = 1$) p.d.f.

The population IQR of the uniform distribution and the exponential distribution were relatively simple to compute because they are each of a closed form. The normal distribution N(μ,σ), however, does not provide the same luxury. By using the "z-score" table, which can be found in most statistics textbooks, it is easy to find both Q_1 and Q_3. We simply find the "z-score" that corresponds to the areas under the normal curve of .25 and .75, respectively. Thus we have

$$Q_3 = .6745\sigma + \mu \tag{28}$$

$$Q_1 = -.6745\sigma + \mu \tag{29}$$

Again, we take the difference of Q_3 and Q_1 which yields,

$$IQR_{\text{norm}} \approx 1.3490\sigma. \tag{30}$$

2.4 The Gamma Distribution

Like the exponential distribution, the gamma distribution is a probability model for waiting times. But we are now interested in the waiting time until the αth occurrence. The waiting time until the αth occurrence in a Poisson process with mean λ can be modeled using a gamma distribution with parameters α and $\beta = 1/\lambda$. The gamma distribution is given by the p.d.f.,

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, \;\; 0 \leq x < \infty \tag{31}$$

The shape of the gamma distribution is skewed right. Also, when $\beta = 2$ and $\alpha = v/2$ where v is a positive integer, the random variable X is said to have a chi-square distribution with v degrees of freedom, which can be denoted by $\chi^2(v)$. Figure 4 is an illustration of a gamma distribution with $\alpha = 2$ and $\beta = 2$ or equivalently a $\chi^2(4)$.

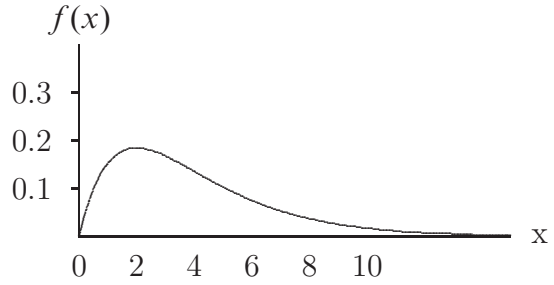

Figure 4: Gamma $\alpha = 2, \beta = 2$ p.d.f.

Applying equations (14), (15) with the gamma p.d.f. (31) we could numerically arrive at the solution. If α is an integer, Q_3 and Q_1 can be found by solving the following equations

$$.75 = 1 - \sum_{x=0}^{\alpha-1} \frac{(Q_3/\theta)^x e^{-Q_3/\theta}}{x!} \tag{32}$$

$$.25 = 1 - \sum_{x=0}^{\alpha-1} \frac{(Q_1/\theta)^x e^{-Q_1/\theta}}{x!}, \tag{33}$$

respectively.

2.5 The Lognormal Distribution

If W is a random variable that is normally distributed (that is, $W \sim N(\mu, \sigma)$), then $X = e^W$ has a lognormal distribution. The lognormal distribution is a skewed right distribution. Modeling using the lognormal enables one to use normal distribution statistics. The lognormal distribution is given by the p.d.f.,

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(lnx-\mu)^2/2\sigma^2}, \ 0 < x < \infty. \tag{34}$$

Figure 5 is an illustration of a lognormal distribution with $\mu = 0$ and $\sigma = 1$.

Using equations (28) and (29) from the normal distribution, it follows that

$$Q_3 = e^{.6745\sigma+\mu} \tag{35}$$

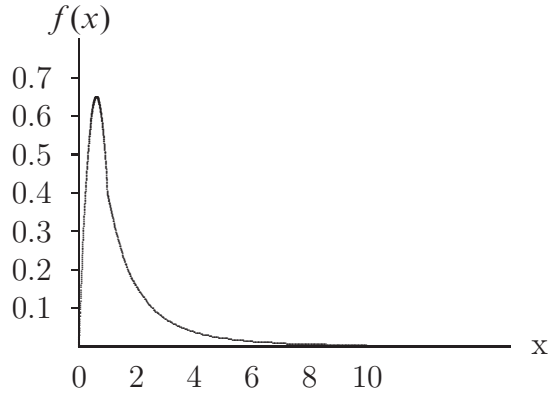

Figure 5: Lognormal $\mu = 0, \sigma = 1$ p.d.f.

and

$$Q_1 = e^{-.6745\sigma+\mu}. \tag{36}$$

Taking the difference between the quartiles in (35) and (36) yields

$$IQR_{\text{logn}} = e^{\mu}\left[e^{.6745\sigma} - e^{-.6745\sigma}\right] = 2e^{\mu}\sinh .6745\sigma \tag{37}$$

where $\sigma > 0$.

2.6 The Weibull Distribution

If E is a random variable that has an exponential distribution with mean β, then $X = E^{1/\gamma}$ has a Weibull distribution with parameters γ and β. The Weibull distribution is an important model in the analysis of the lifetime of a product. The Weibull distribution is given by the p.d.f.,

$$f(x) = \frac{\gamma}{\beta}x^{\gamma-1}e^{-x^{\gamma}/\beta},\ 0 < x < \infty. \tag{38}$$

The Weibull distribution is a skewed distribution. Figure 6 is an illustration of a Weibull distribution with $\gamma = 10$ and $\beta = .5$.

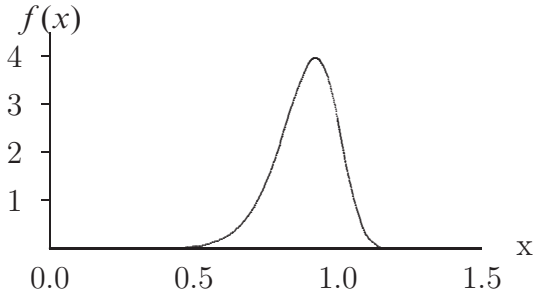

Figure 6: Weibull $\gamma = 10, \beta = .5$ p.d.f.

It is not difficult to find Q_1, Q_3, and the IQR for the Weibull distribution. By definition,

$$\begin{aligned} .25 &= P(X \leq Q_1) \\ &= P(E^{1/\gamma} \leq Q_1) \\ &= P(E \leq Q_1^{\gamma}). \end{aligned} \tag{39}$$

Since we know from equation (25) that Q_1 of an exponential distribution with mean β is $-\beta \ln(.75)$, it is readily seen that

$$Q_1 = [-\beta \ln(.75)]^{1/\gamma} \tag{40}$$

for the Weibull distribution with parameters γ and β. Similarly, it follows that

$$Q_3 = [-\beta \ln(.25)]^{1/\gamma}. \tag{41}$$

Hence,

$$\begin{aligned} IQR_{\text{weib}} &= [-\beta \ln(.25)]^{1/\gamma} - [-\beta \ln(.75)]^{1/\gamma} \\ &= \beta^{1/\gamma}\{[\ln(4)]^{1/\gamma} - [\ln(4/3)]^{1/\gamma}\}. \end{aligned} \tag{42}$$

3 THE EXPECTED VALUE OF THE SAMPLE IQR

In this chapter, we will compute the expected value of the sample IQR when a random sample is taken from the uniform, exponential, and chi-square distributions. The expected value of a continuous random variable, X, is defined as,

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{43}$$

given that

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty. \tag{44}$$

Now, let $X_1, X_2, \ldots, X_n$ be independent identically distributed (iid) continuous random variables with distribution function $F(y)$ and density function $f(y)$. Let Y_1 be the smallest of the X_i's; let Y_2 denote the second smallest of the X_i's, $\cdots$, and let Y_n be the largest of the X_i's. The random variables $Y_1 < Y_2 < \cdots < Y_n$ are called the "order statistics" of the sample. If Y_k denotes the k^{th} order statistic, then the density function of Y_k is given by

$$g_{(k)}(y_k) = \frac{n!}{(k-1)!(n-k)!}[F(y_k)]^{k-1}[1 - F(y_k)]^{n-k} f(y_k), \quad -\infty < y_k < \infty. \tag{45}$$

We are specifically interested in the order statistics that correspond to q_3 and q_1. For example, if $n = 11$ then Y_3 and Y_9 are the order statistics that correspond to q_1 and q_3, respectively.

3.1 Uniform Distribution

In order to find the s-order statistic we will substitute for $f(x)$ and $F(x)$ into equation (45), where $f(x)$ is given in equation (17), and $F(x)$ is found by,

$$\begin{aligned} F(x) &= \int_{\theta_1}^{x} \frac{1}{\theta_2 - \theta_1} dt \\ &= \frac{x - \theta_1}{\theta_2 - \theta_1}. \end{aligned} \tag{46}$$

Since we are concerned with the s-order statistic we let $x = y_s$,

$$\begin{aligned} g_{(s)}(y_s) = \frac{n!}{(s-1)!(n-s)!} \left[\frac{y_s - \theta_1}{\theta_2 - \theta_1}\right]^{s-1} \left[1 - \frac{y_s - \theta_1}{\theta_2 - \theta_1}\right]^{n-s} \\ \times \left(\frac{1}{\theta_2 - \theta_1}\right), \ \theta_1 < y_s < \theta_2 \end{aligned} \tag{47}$$

For the purpose of this paper, we will consider the uniform distribution only from zero to one (i.e. unif[0,1]). Thus, by substituting into equation (43) and letting $\theta_1 = 0$ and $\theta_2 = 1$ we have,

$$E(Y_s) = \int_0^1 y_s \frac{n!}{(s-1)!(n-s)!} y_s^{s-1} [1 - y_s]^{n-s} \, dy_s. \tag{48}$$

Now that we can find the expected values for a specific order statistic, we need only take the difference of the order statistics that correspond to the third and first quartiles to find the expected value of the sample interquartile range (denoted $\widehat{IQR}$). Now, let q_1 correspond to the order statistic Y_r and let q_3 correspond to the order statistic Y_s. Applying equation (48) we have

$$\begin{aligned} E(\widehat{IQR}) &= E(Y_s) - E(Y_r) \\ &= \int_0^1 y_s \frac{n!}{(s-1)!(n-s)!} y_s^{s-1} [1 - y_s]^{n-s} \, dy_s \\ &- \int_0^1 y_r \frac{n!}{(r-1)!(n-r)!} y_r^{r-1} [1 - y_r]^{n-r} \, dy_r. \end{aligned} \tag{49}$$

With the aid of Maple, equation (49) simplifies to

$$E(\widehat{IQR}) = \frac{s - r}{1 + n} \tag{50}$$

where r is the rth order statistic and s is the sth order statistic. By simply knowing the sample size, we can now easily calculate the expected value of $\widehat{IQR}$ when the underlying distribution is a unif[0,1]. In fact, we note that $E(\widehat{IQR}) = .5$. For example, if $n = 11$ we know that $q_3 = Y_9$ and $q_1 = Y_3$. Applying equation (50) with $n = 11, s = 9$ and $r = 3$ yields .5. We further note that the sample interquartile range is an unbiased estimator of the population interquartile range.

3.2 Exponential Distribution

In this section, we find $E(\widehat{IQR})$ when the underlying distribution is an exponential distribution. We substitute $f(x)$ and $F(x)$ into equation (45), where $f(x)$ is given in equation (22) and $F(x)$ is found by,

$$\begin{aligned} F(x) &= \int_0^x \frac{1}{\beta} e^{-t/\beta} dt \\ &= 1 - e^{x/\beta}. \end{aligned} \tag{51}$$

Thus the p.d.f. of Y_s is

$$g_{(s)}(y_s) = \frac{n!}{(s-1)!(n-s)!} \left[1 - e^{-(y_s/\beta)}\right]^{s-1} \left[e^{-(y_s/\beta)}\right]^{n-s} \left(\frac{1}{\beta} e^{-(y_s/\beta)}\right). \tag{52}$$

For the purpose of this paper we are only going to consider exp(1). That is, using $\beta = 1$ in equation (52) we have the expected value of Y_s

$$E(Y_s) = \int_0^\infty y_s \frac{n!}{(s-1)!(n-s)!} [1 - e^{-y_s}]^{s-1} [e^{-y_s}]^{n-s+1} \, dy_s. \tag{53}$$

Because there are so many variables in this equation, the integral does not simplify very easily. Unfortunately, the expected value of $\widehat{IQR}$ is also very difficult to simplify.

Thus, we have,

$$\begin{aligned} E(\widehat{IQR}) = &\int_0^\infty y_s \frac{n!}{(s-1)!(n-s)!} [1-e^{-y_s}]^{s-1} [e^{-y_s}]^{n-s+1} \, dy_s \\ &- \int_0^\infty y_r \frac{n!}{(r-1)!(n-r)!} [1-e^{-y_r}]^{r-1} [e^{-y_r}]^{n-r+1} \, dy_r. \end{aligned} \tag{54}$$

The expected value of $\widehat{IQR}$ for some specific sample sizes are given in Table 2. We note that the sample interquartile range is a biased estimator of the population interquartile range. Table 2 reveals that $E(\widehat{IQR})$ is greater than 1.0986 which was found in equation (26).

3.3 Chi-Square Distribution

The chi-square distribution is the last distribution for which we calculate the expected value of $\widehat{IQR}$. For simplicity, we will consider a chi-square distribution with four degrees of freedom, denoted by χ^2_4. The p.d.f. for the χ^2_4 is given by

$$f(x) = \frac{1}{4} x e^{-x/2}, \; 0 < x < \infty \tag{55}$$

and the distribution function $F(x)$ is

$$\begin{aligned} F(x) &= \int_0^x \frac{1}{4} t e^{-t/2} dt \\ &= 1 - e^{-x/2} - \frac{1}{2} x e^{-x/2}. \end{aligned} \tag{56}$$

It follows that the expected value of $\widehat{IQR}$, using the appropriate r and s values that correspond to the first and third quartiles, is based on

$$\begin{aligned} E(Y_k) = &\int_0^\infty y_k \frac{n!}{(k-1)!(n-k)!} \left[1 - e^{-y_k/2} - \frac{1}{2} y_k e^{-y_k/2} \right]^{k-1} \\ &\cdot \left[1 - \left(1 - e^{-y_k/2} - \frac{1}{2} y_k e^{-y_k/2} \right) \right]^{n-k} \left(\frac{1}{4} y_s e^{-y_k/2} \right) dy_k. \end{aligned} \tag{57}$$

As was the case with the exponential distribution, the expected value of Y_k is tedious to calculate, even with the aid of software such as Maple. Since we are unable to provide a closed form for $E(\widehat{IQR})$, computations using Maple for specific sample sizes are given in Table 2. We note that the sample interquartile range is a biased estimator of the population interquartile range which is equal to 3.4627.

4 THE VARIANCE OF THE SAMPLE IQR

Chapter 3 discussed the computation of the expectation of the sample interquartile range for the uniform, exponential, and chi-square distributions. In this chapter, the variance of the sample interquartile range is presented. In order to calculate the variance of $\widehat{IQR}$, we must first find the variance of the order statistics corresponding to q_3 and q_1. We will continue to use the rth order statistic for q_1 and the sth order statistic for q_3. The variance of a random variable, Y, is defined by

$$V(Y) = E(Y^2) - [E(Y)]^2. \tag{58}$$

The variance of $\widehat{IQR}$, however, is not as simple to calculate as the expected value of $\widehat{IQR}$. In general, the variance of Y_k can be very tedious to find and sometimes no closed form exists. The problem of finding the variance of the difference between two order statistics gets even more complicated. We can not simply take the difference of $V(Y_s)$ and $V(Y_r)$. The covariance (Cov) between Y_s and Y_r must also be considered in this calculation [2]. Examples of the computational difficulties in computing the variance of the sample median can be found in the papers by Maritz and Jarrett[6] and Price and Bonett [9]. The variance of the sample interquartile range is defined by

$$V(\widehat{IQR}) = V(Y_s) + V(Y_r) - 2 \cdot \mathrm{Cov}(Y_s, Y_r) \tag{59}$$

where the covariance is given by

$$\mathrm{Cov}(Y_r, Y_s) = E(Y_r Y_s) - E(Y_r)E(Y_s). \tag{60}$$

In next three sections, the variance of the sample interquartile range is given for the uniform, exponential, and the chi-square distributions.

4.1 Uniform Distribution

We begin by finding variance of Y_s when the underlying distribution is the unif[0,1], i.e.,

$$\begin{aligned} V(Y_s) = &\int_0^1 y_s^2 \frac{n!}{(s-1)!(n-s)!}(y_s)^{s-1}\,[1-y_s]^{n-s}\,dy \\ &- \left(\int_0^1 y_s \frac{n!}{(s-1)!(n-s)!}(y_s)^{s-1}\,[1-y_s]^{n-s}\,dy\right)^2. \end{aligned} \tag{61}$$

After we integrate and simplify we find the following for the variance of the sth order statistic

$$V(Y_s) = -\frac{s(-n+s-1)}{(n+2)(n+1)^2} \tag{62}$$

and similarly,

$$V(Y_r) = -\frac{r(-n+r-1)}{(n+2)(n+1)^2}. \tag{63}$$

So, now that we know the values of $V(Y_s)$ and $V(Y_r)$, we need only to find the $\mathrm{Cov}(Y_r, Y_s)$ in order to solve equation (59). In Chapter 3, we discussed, in detail, how to find $E(Y_r)$ and $E(Y_s)$. The $E(Y_rY_s)$, however, is a little more difficult. It is found by

$$E(Y_rY_s) = \int_{-\infty}^{\infty}\int_{-\infty}^{y_s} y_r \cdot y_s f(y_r, y_s)\,dy_r\,dy_s \tag{64}$$

where,

$$\begin{aligned} f(y_r, y_s) = &\frac{n!}{(r-1)!(s-r-1)!(n-s)!}\,[F(y_r)]^{(r-1)}\,[F(y_s)-F(y_r)]^{(s-r-1)} \\ &\times [1-F(y_s)]^{(n-s)}\,f(y_r)f(y_s), -\infty < y_r < y_s < \infty. \end{aligned} \tag{65}$$

Now, by substituting equation (65) for $f(y_r, y_s)$ in equation (64), we can finally calculate the covariance which, in turn, will allow us to calculate the variance of the

$\widehat{IQR}$ by substituting into equation (59). Since we already know the variance of Y_s and Y_r from equation (62) we need only find the covariance of Y_s, Y_r. As the equation is now, it is very difficult to solve this integral. If, however, we use Maple and substitute the specific functions that pertain to the function unif[0,1], the covariance is found to be

$$\text{Cov}(Y_s, Y_r) = -\frac{r(-n+s-1)}{(n+1)^2(n+2)}. \tag{66}$$

We can now substitute into equation 59 to find the variance of the sample interquartile range for the unif[0,1],

$$V(\widehat{IQR}) = \frac{s(n-s+1)}{(n+2)(n+1)^2} + \frac{r(n-r+1)}{(n+2)(n+1)^2} + \frac{2r(-n+s-1)}{(n+1)^2(n+2)}. \tag{67}$$

Again, the unif[0,1] gives a nice solution. From this equation, we see that we need only know the r, s, and n values to calculate the variance of $\widehat{IQR}$ for the unif[0,1]. For example, if a random sample of size $n = 39$ is drawn from a unif[0,1] distribution, then $r = 10$, $s = 30$ and substituting these values into equation (67) yields

$$\begin{aligned} V(\widehat{IQR}) = &\frac{30(39-30+1)}{(39+2)(39+1)^2} + \frac{10(39-10+1)}{(39+2)(39+1)^2} \\ &+ \frac{2(10)(-39+30-1)}{(39+1)^2(39+2)} = .0061. \end{aligned} \tag{68}$$

Table 2 shows the variance of the sample interquartile range for several different values of n and compares these values to estimates of this variance using a simulation method called "bootstrapping", which will be discussed in Chapter 5.

4.2 Exponential

Of course, the steps for finding the variance of $\widehat{IQR}$ when the underlying distribution follows the exponential with β_1 are similar to that of the unif[0,1]. After

substituting into equation (58) we have,

$$\begin{aligned} V(Y_s) = &\int_0^\infty y_s^2 \frac{n!}{(s-1)!(n-s)!} [1-e^{-y_s}]^{s-1} [e^{-y_s}]^{n-s} (e^{-y_s})\, dy_s \\ &- \left[\int_0^\infty y_s \frac{n!}{(s-1)!(n-s)!} [1-e^{-y_s}]^{s-1} [e^{-y_s}]^{n-s} (e^{-y_s})\, dy_s\right]^2. \end{aligned} \tag{69}$$

Unfortunately, even with Maple, it is very difficult to simplify the variance. The same is true for the covariance,

$$\begin{aligned} \mathrm{Cov}(Y_s, Y_r) = &\int_0^\infty \int_0^{y_s} y_r y_s \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \\ &(1-e^{-y_r})^{(r-1)} (-e^{-y_s})^{(s-r-1)} (e^{-y_s})^{(n-s)} e^{-y_r} e^{-y_s} dy_r dy_s \\ &- \left(\int_0^\infty y_s \frac{n!}{(s-1)!(n-s)!} [1-e^{-y_s}]^{s-1} [e^{-y_s}]^{n-s} (e^{-y_s})\, dy_s \right. \\ &\left. \times \int_0^\infty y_r \frac{n!}{(r-1)!(n-r)!} [1-e^{-y_r}]^{r-1} [e^{-y_r}]^{n-r} (e^{-y_r})\, dy_r \right). \end{aligned} \tag{70}$$

Bringing equations (69) and (70) together, we find the variance of the sample interquartile range can be written as

$$\begin{aligned} V(\widehat{IQR}) = &\int_0^\infty y_s^2 \frac{n!}{(s-1)!(n-s)!} [1-e^{-y_s}]^{s-1} [e^{-y_s}]^{n-s} (e^{-y_s})\, dy_s \\ &- \left[\int_0^\infty y_s \frac{n!}{(s-1)!(n-s)!} [1-e^{-y_s}]^{s-1} [e^{-y_s}]^{n-s} (e^{-y_s})\, dy_s\right]^2 \\ &+ \int_0^\infty y_r^2 \frac{n!}{(r-1)!(n-r)!} [1-e^{-y_r}]^{r-1} [e^{-y_r}]^{n-r} (e^{-y_r})\, dy_r \\ &- \left[\int_0^\infty y_r \frac{n!}{(r-1)!(n-r)!} [1-e^{-y_r}]^{r-1} [e^{-y_r}]^{n-r} (e^{-y_r})\, dy_r\right]^2 \\ &- 2\left[\int_0^\infty \int_0^{y_s} y_r y_s \frac{n!}{(r-1)!(s-r-1)!(n-s)!} (1-e^{-y_r})^{(r-1)} \right. \\ &\times (-e^{-y_s})^{(s-r-1)} (e^{-y_s})^{(n-s)} e^{-y_r} e^{-y_s} dy_r dy_s \\ &- \left(\int_0^\infty y_s \frac{n!}{(s-1)!(n-s)!} [1-e^{-y_s}]^{s-1} [e^{-y_s}]^{n-s} (e^{-y_s})\, dy_s \right. \\ &\left.\left. \times \int_0^\infty y_r \frac{n!}{(r-1)!(n-r)!} [1-e^{-y_r}]^{r-1} [e^{-y_r}]^{n-r} (e^{-y_r})\, dy_r \right)\right]. \end{aligned} \tag{71}$$

Even though this is a very difficult equation to solve because of the many variables, we can substitute for r, s, and n in our sample to get the theoretical variance of $\widehat{IQR}$. We show these values in Table 2.

4.3 Chi-Square (χ^2) Distribution

The procedure for calculating the variance of $\widehat{IQR}$ for the χ^2_4 distribution is exactly the same as was done for the uniform and the exponential distributions. Substituting into equation (58), we find

$$\begin{aligned} V(Y_s) = &\int_0^\infty y_s^2 \frac{n!}{(s-1)!(n-s)!} \left[-e^{-y_s/2} - \frac{1}{2} y_s e^{-y_s/2} + 1\right]^{s-1} \\ &\cdot \left[1 - \left(-e^{-y_s/2} - \frac{1}{2} y_s e^{-y_s/2} + 1\right)\right]^{n-s} \left(\frac{1}{4} y_s e^{-y_s/2}\right) dy_s \\ &- \left[\int_0^\infty y_s \frac{n!}{(s-1)!(n-s)!} \left[-e^{-y_s/2} - \frac{1}{2} y_s e^{-y_s/2} + 1\right]^{s-1}\right. \\ &\left. \cdot \left[1 - \left(-e^{-y_s/2} - \frac{1}{2} y_s e^{-y_s/2} + 1\right)\right]^{n-s} \left(\frac{1}{4} y_s e^{-y_s/2}\right) dy_s\right]^2 \end{aligned} \tag{72}$$

As was the case with the exponential distribution, the variance of the sample interquartile range for the χ^2_4 is difficult to calculate and does not simplify very well. Fortunately, with the aid of Maple, we can calculate the variance of $\widehat{IQR}$ for specific values of n. We give the values of the variance for some values of n in table 2, specifically when $n = 4m - 1$ and m is an integer.

5 BOOTSTRAPPING

In the two previous chapters, we looked at some of the properties of the sample interquartile range. In this chapter, we are interested in estimating the population interquartile range, the expected value, and variance of the sample interquartile range. Since it is not always feasible that the underlying distribution be known, we consider a method called "bootstrapping". The bootstrap was introduced in 1979 as a method for estimating the standard error of $\widehat{\theta}$. [3] In the context of this paper, the parameter of interest is the population interquartile range which was discussed in Chapter 2. Efron defines the bootstrap to be a random sample of size n drawn from $\widehat{F}$, where $\widehat{F}$ is the empirical distribution function. The empirical distribution function is the discrete distribution that puts probability $1/n$ on each value $x_i, i = 1, 2, ..., n$. Where the $x_{i'}s$ are random observations from some probability distribution F, such that,

$$F \rightarrow (x_1, x_2, ..., x_n). \tag{73}$$

Put simply, $\widehat{F}$ assigns to set A, in the sample space of x, its empirical probability,

$$\widehat{\text{Prob}}\{A\} = \#\{x_i \in A\}/n \tag{74}$$

the proportion of the observed sample $x = (x_1, x_2, ...x_n)$ occurring in A. So, by drawing from the empirical distribution, $\widehat{F}$, say $x^* = (x_1^*, x_2^*, ..., x_n^*)$,

$$\widehat{F} \rightarrow (x_1^*, x_2^*, ..., x_n^*) \tag{75}$$

The star notation, *, is used to signify that x^* is a "randomized", or "resampled" version of x and not the actual data set x. Basically, the procedure for the bootstrap is to obtain B independent data sets $(x_1^*, x_2^*, ..., x_n^*)$ by sampling n observations, with

replacement, from the population of n objects $(x_1, x_2, ...x_n)$.[3] To aid in illustrating how the bootstrap works, consider the following, very basic, example.

Let $x = (0, 1, 2, 3, 4, 5, 6, 7, 8, 9)$ be the population of ten objects. Following the bootstrap procedure, sample 10 observations with replacement from the given population. By using any random number generating technique, one possibility for x^* might be $x^* = (4, 3, 5, 3, 6, 1, 6, 7, 1, 5)$. Because several of these samples must be made, the first x^* can be notated as x^{*1} the second sample, x^{*2} and so on to the Bth sample, x^{*B}. Thus, if $B = 5$, the following is one possibility for the five bootstrapped samples

$$\begin{aligned} x^{*1} &= (4, 3, 5, 3, 6, 1, 6, 7, 1, 5) \\ x^{*2} &= (0, 8, 3, 6, 7, 2, 9, 5, 0, 7) \\ x^{*3} &= (9, 9, 2, 2, 7, 1, 1, 1, 9, 2) \\ x^{*4} &= (6, 9, 2, 0, 9, 7, 7, 2, 6, 9) \\ x^{*5} &= (5, 5, 0, 1, 7, 8, 8, 3, 4, 5) \end{aligned} \tag{76}$$

Once the bootstrap samples have been taken, any statistic of interest can be calculated from each sample. By the central limit theorem, as $n \to \infty$, the bootstrap histogram will become bell-shaped, however, for a relatively small number of bootstrap samples the histogram may not be normal. Clearly a statistic based on five samples (as in the hypothetical sample above) would probably not provide a normal histogram. If it were possible to take $B = \infty$ bootstrap samples, the calculation of any statistic from this sample would be an exact estimate of the parameter. Obviously, taking an infinite sample is impossible. So, how many bootstrap samples are reasonable to take

to accurately estimate the desired parameter? According to Efron, if one is estimating the standard error, se, 50 to 200 samples are plenty. However, when calculating the confidence intervals, 1000 or 2000 samples would be more appropriate. Drawing 1000 samples by hand would be impractical, not to mention time consuming. Fortunately, with the technology available today, computer simulations can quickly draw 1000 or 2000 bootstrap samples. The confidence intervals calculated in this paper were done using a computer program called MATLAB. The MATLAB code used to generate these bootstrap samples is given in the appendix.

Table 2 shows the results from the bootstrap estimates "computed" along side of the theoretical expected value and variance of the sample interquartile range for the uniform, exponential, and chi-square distributions. The theoretical expected value and variance are in the columns headed E($\widehat{\text{IQR}}$) and V($\widehat{\text{IQR}}$), respectively. The bootstrap estimates of the expected value and variance are in the columns headed E[E(IQR*)] and E[V(IQR*)], respectively. Based on the observations in Table 2, it would appear that the bootstrap estimate is fairly close to the expected value of the sample interquartile range. The bootstrap variance tends to be over estimating the true variance. It would also appear that the bootstrap estimates converge to the theoretical values as n becomes increasingly large. In the next section, we look at the accuracy of a bootstrap confidence interval for the population interquartile range.

5.1 Percentile Method

One of the nice things about bootstrapping the confidence interval is that it can generally be used without the knowledge of the underlying distribution from which

observations are drawn. Efron mentions several different methods for calculating the confidence interval of a parameter once the bootstrap samples have been generated. For this paper a relatively simple method for estimating the confidence interval called the "percentile method" is used. In general for the usual plug in estimate, $\hat{\theta}$, of the parameter θ, let $\widehat{se}$ be the estimated standard error for $\hat{\theta}$. Then we have the standard normal confidence interval, $[\hat{\theta} - z_{1-\alpha/2} \cdot \widehat{se}, \hat{\theta} + z_{1-\alpha/2} \cdot \widehat{se}]$, where $P(Z > z_{1-\alpha/2}) = 1 - \alpha/2$. So for the bootstrap, if $\hat{\theta}^*$ is a random variable drawn from the distribution $N(\hat{\theta}, \widehat{se}^2)$ then the confidence interval is

$$[\hat{\theta}^{*(\alpha/2)}, \hat{\theta}^{*(1-\alpha/2)}]. \tag{77}$$

Equation (77) refers to the ideal bootstrap situation: an infinite number of bootstrap replications. A finite number, $\boldsymbol{B}$, of independent bootstrap replications will be drawn instead. The statistic of interest, in this case the $\widehat{\text{IQR}}$, is calculated for each independent bootstrapped data set. Then $\hat{\theta}_{\boldsymbol{B}}^{*(\alpha)}$ is the $100 \cdot \alpha$th empirical percentile of the $\hat{\theta}^*(b)$ values, which is the $\boldsymbol{B} \cdot \alpha$th value in the ordered list of the $\boldsymbol{B}$ replications of $\hat{\theta}^*$. So, the confidence intervals for this paper were calculated using 2000 replications, that is, $\boldsymbol{B} = 2000$. Three alpha levels were chosen to calculate the confidence intervals, $\alpha = 0.1$, $\alpha = 0.05$, and $\alpha = 0.01$. For example, if we are interested in computing a 90% confidence interval then $\alpha = 0.1$. So, after the $\widehat{\text{IQR}}$ has been calculated for each of the $\boldsymbol{B}$ replications and ordered, the lower bound for the confidence interval would be the $2000 \cdot (0.1/2)$ or 100th order statistic. Similarly, the upper bound would be the $2000 \cdot (1 - .1/2) + 1$ or 1901th order statistic.

A simulated study was conducted and the results are summarized in Tables 3 - 8. Each row of a table represents the empirical coverage of the percentile method based

on 1000 simulated variates from the given distribution. That is, the empirical coverage that the "population IQR" lies between the lower and upper bounds of the bootstrap confidence interval. Two-thousand bootstrap samples were employed. Notice that the simulated "coverage result" is typically higher than the "confidence coefficient" (i.e. $1-\alpha$). Ideally, the confidence coefficient and the coverage result would be equal. Since the coverage result is typically higher than the given confidence coefficient, this is an indication that the percentile confidence interval gives a slightly conservative estimate of the population IQR.

6 CONCLUSION

The interquartile range is a valuable tool for describing the spread of a given set of data. It is typically used when the shape of the distribution is skewed or outliers are present. The interquartile range is easy to compute for samples of size n, as is the population interquartile range for continuous probability distributions. The expected value and the variance of $\widehat{\text{IQR}}$, however, are not easily calculated. Fortunately, software programs, such as Maple, have been developed to aid in these calculations.

We have shown that $\widehat{\text{IQR}}$ is an unbiased estimator of the population interquartile range when a random sample is taken from the uniform distribution. However, the expected value of $\widehat{\text{IQR}}$ did not equal the population interquartile range for the exponential and chi-square distributions. Thus $\widehat{\text{IQR}}$ is a biased estimator of the population interquartile range for these skewed distributions. We hypothesize that this may be true for other skewed distributions.

The population IQR, expected value, and variance of the sample IQR each required knowledge of the "underlying" distribution. In an attempt to find an estimation method that did not require a knowledge of the distribution, we chose to use a simulation method called the bootstrap. Due to the large amount of "resampling" involved in the bootstrap method, computer software was again required. We chose to use MATLAB. The bootstrap mean was shown to estimate the expected value of $\widehat{\text{IQR}}$ relatively effectively. The bootstrap variance tended to over estimate the true variance of $\widehat{\text{IQR}}$. We also investigated the usefulness of a bootstrapped confidence interval for the IQR. The percentile method tended to be conservative but the coverage

converged to the "confidence coefficient" as n became increasingly large.

Table 2: Bootstrap and Theoretical Mean and Variance of the Sample IQR.

Uniform Distribution						
n	r	s	E($\widehat{IQR}$)	E[E(IQR*)]	V($\widehat{IQR}$)	E[V(IQR*)]
11	3	9	0.5000	0.4573	0.0192	0.0252
15	4	12	0.5000	0.4696	0.0147	0.0193
19	5	15	0.5000	0.4606	0.0199	0.0154
23	6	18	0.5000	0.4787	0.0100	0.0128
27	7	21	0.5000	0.4836	0.0086	0.0109
31	8	24	0.5000	0.4828	0.0076	0.0095
35	9	27	0.5000	0.4855	0.0068	0.0084
39	10	30	0.5000	0.4875	0.0061	0.0075
43	11	33	0.5000	0.4908	0.0056	0.0068
47	12	36	0.5000	0.4885	0.0051	0.0062
51	13	39	0.5000	0.4906	0.0047	0.0057
99	25	75	0.5000	0.4952	0.0025	0.0029
Exponential Distribution						
11	3	9	1.2179	1.2213	0.2774	0.4715
15	4	12	1.1865	1.1818	0.1969	0.3103
19	5	15	1.1682	1.1643	0.1524	0.2205
23	6	18	1.1562	1.1604	0.1242	0.1741
27	7	21	1.1477	1.1504	0.1048	0.1468
31	8	24	1.1414	1.1439	0.0906	0.1222
35	9	27	1.1366	1.1382	0.0798	0.1059
39	10	30	1.1327	1.1256	0.0713	0.0929
43	11	33	1.1295	1.1340	0.0644	0.0838
47	12	36	1.1269	1.1279	0.0587	0.0742
51	13	39	1.1247	1.1228	0.0540	0.0673
99	25	75	1.1121	1.1123	0.0274	0.0322
Chi-Square Distribution						
11	3	9	3.7801	3.7622	2.0312	3.2662
15	4	12	3.6976	3.6628	1.4619	2.2195
19	5	15	3.6491	3.6479	1.1408	1.6453
23	6	18	3.6172	3.5837	0.9350	1.2895
27	7	21	3.5846	3.5550	0.7920	1.0573
31	8	24	3.5778	3.5584	0.6869	0.9106
35	9	27	3.5647	3.5581	0.6063	0.7996
39	10	30	3.5544	3.5365	0.5427	0.7083
43	11	33	3.5459	3.5377	0.4911	0.6228
47	12	36	3.5389	3.5220	0.4485	0.5546

Table 3: Uniform Distribution 2000 Bootstraps; 1000 Simulations

Uniform Distribution		
n	$1-\alpha$	Coverage Result
10	.90	0.9915
	.95	0.9975
	.99	1.0000
25	.90	0.9760
	.95	0.9920
	.99	1.0000
50	.90	0.9570
	.95	0.9865
	.99	0.9980
75	.90	0.9390
	.95	0.9800
	.99	0.9980
100	.90	0.9525
	.95	0.9790
	.99	0.9970
133	.90	0.9460
	.95	0.9765
	.99	0.9955
166	.90	0.9430
	.95	0.9735
	.99	0.9965
199	.90	0.9250
	.95	0.9735
	.99	0.9980

Table 4: Exponential Distribution 2000 Bootstraps; 1000 Simulations

Exponential Distribution		
n	$1-\alpha$	Coverage Result
10	.90	0.9435
	.95	0.9760
	.99	0.9870
25	.90	0.9360
	.95	0.9690
	.99	0.9990
50	.90	0.9130
	.95	0.9610
	.99	0.9955
75	.90	0.9150
	.95	0.9655
	.99	0.9945
100	.90	0.9125
	.95	0.9645
	.99	0.9905
133	.90	0.9070
	.95	0.9640
	.99	0.9930
166	.90	0.9150
	.95	0.9595
	.99	0.9925
199	.90	0.9195
	.95	0.9580
	.99	0.9910

Table 5: Chi-Square Distribution 2000 Bootstraps; 1000 Simulations

Chi-Square (χ^2) Distribution		
n	$1-\alpha$	Coverage Result
10	.90	0.9540
	.95	0.9855
	.99	0.9980
25	.90	0.9390
	.95	0.9855
	.99	0.9995
50	.90	0.9370
	.95	0.9670
	.99	0.9965
75	.90	0.9310
	.95	0.9700
	.99	0.9975
100	.90	0.9310
	.95	0.9640
	.99	0.9965
133	.90	0.9320
	.95	0.9675
	.99	0.9933
166	.90	0.9205
	.95	0.9640
	.99	0.9950
199	.90	0.9150
	.95	0.9595
	.99	0.9955

Table 6: Lognormal Distribution 2000 Bootstraps; 1000 Simulations

Lognormal Distribution		
n	$1-\alpha$	Coverage Result
10	.90	0.9315
	.95	0.9725
	.99	0.9890
25	.90	0.9150
	.95	0.9550
	.99	0.9930
50	.90	0.9130
	.95	0.9660
	.99	0.9930
75	.90	0.9190
	.95	0.9635
	.99	0.9935
100	.90	0.9015
	.95	0.9490
	.99	0.9930
133	.90	0.9255
	.95	0.9540
	.99	0.9905
166	.90	0.9060
	.95	0.9490
	.99	0.9930
199	.90	0.9075
	.95	0.9510
	.99	0.9905

Table 7: Cauchy Distribution 2000 Bootstraps; 1000 Simulations

Cauchy Distribution		
n	$1-\alpha$	Coverage Result
10	.90	0.7950
	.95	0.9100
	.99	0.9895
25	.90	0.8780
	.95	0.9560
	.99	0.9915
50	.90	0.9070
	.95	0.9500
	.99	0.9925
75	.90	0.9380
	.95	0.9685
	.99	0.9970
100	.90	0.9300
	.95	0.9705
	.99	0.9970
133	.90	0.9190
	.95	0.9625
	.99	0.9930
166	.90	0.9170
	.95	0.9595
	.99	0.9915
199	.90	0.9205
	.95	0.9665
	.99	0.9970

Table 8: Laplace Distribution 2000 Bootstraps; 1000 Simulations

Laplace (Double Exp(1)) Distribution		
n	$1-\alpha$	Coverage Result
10	.90	0.8570
	.95	0.9645
	.99	0.9980
25	.90	0.9195
	.95	0.9660
	.99	0.9975
50	.90	0.9200
	.95	0.9610
	.99	0.9905
75	.90	0.9315
	.95	0.9730
	.99	0.9955
100	.90	0.9315
	.95	0.9695
	.99	0.9950
133	.90	0.9320
	.95	0.9650
	.99	0.9950
166	.90	0.9225
	.95	0.9630
	.99	0.9930
199	.90	0.9290
	.95	0.9665
	.99	0.9960

BIBLIOGRAPHY

[1] Agresti, Alan and Christine A. Franklin (2005) *STATISTICS: The Art and Science of Learning from Data*, Pearson Prentice Hall.

[2] David, Herbert A. (1969) *Order Statistics 2nd edition*, John Wiley & Sons.

[3] Efron, B. and R. J. Tibshirani (1993) *An Introduction to the Bootstrap*, London: Chapman and Hall.

[4] Hettmansperger, T.P and Sheather, S.J. (1986) Confidence Intervals Based on Interpolated Order Statistics. *Statist. Probab. Lett.*, **4**, 75 - 79.

[5] Hogg, Robert V. and Elliot A. Tanis (2001) *Probability and Statistical Inferences, 6th ed.* Prentice Hall.

[6] Maritz, J.S. and R.G. Jarrett (1978) A Note on Estimating the Variance of the Sample Median. *J. Amer. Statist. Assoc.*, **73**, 194 - 196.

[7] Martinez,Wendy L. and Angel R. Martinez. (2002) *Computational Statistics Handbook with MATLAB.*

[8] Moore, David S. (2000) *The Basic Practice of Statistics, 2nd ed.* W.H. Freeman and Company.

[9] Price, Robert M. and Douglas G. Bonett (2001) Estimating the Variance of the Sample Median. *J. Statist. Comput. Simul.*,**68**, 295 - 305.

[10] Wackerly, Dennis, Richard L. Scheaffer, and William Mendenhall (1996) *Mathematical Statistics with Applications, 5th ed.* Duxbury Press.

APPENDICES

.1 Matlab Code Exp(1) and Chi2(4)

```
function[mu,variance,cov] = boot_iqrs1(dist,parm1,n,alpha) tic;

r=floor((n+1)/4);

s = floor(3*(n+1)/4);

ratio1 = mod(n+1,4)/4;

ratio2 = mod(3*(n+1),4)/4;

delta = icdf (dist,.75,parm1) -icdf(dist,.25,parm1);

iqr_pop = []; mn_boot = [];

var_boot = [];

lb_iqr=[]; ub_iqr=[];


for i = 1:2000


x=random(dist,parm1,n,1);

 x=sort(x);

 q1 = (x(r) + ratio1*(x(r+1)-x(r)));

 q3 = (x(s) + ratio2*(x(s+1)-x(s)));

 iqrange=q3-q1;

 u = random('unid',n,n,2000);

 boot_samples = x(u);

 sort_boot = sort(boot_samples);
```

```
 iqr_boot = (sort_boot(s,:)+ratio2*(sort_boot(s+1,:)
 -sort_boot(s,:))-sort_boot(r,:)+ratio1*(sort_boot(r+1,:)-sort_boot(r,:)));
 iqr_sort = sort(iqr_boot);
 lb = iqr_sort(2000*alpha/2);
 ub = iqr_sort(2000*(1-alpha/2) + 1);
 lb_iqr = [lb_iqr;lb];
 ub_iqr = [ub_iqr;ub];
 meaniqr = mean(iqr_boot);
 variqr = var(iqr_boot);
  iqr_pop = [iqr_pop;iqrange];
  mn_boot = [mn_boot; meaniqr];
  var_boot = [var_boot; variqr];



end
mu=mean(mn_boot);
variance=mean(var_boot);
cov = sum((lb_iqr
<= delta & delta <= ub_iqr))/2000;


toc;
```

.2 Matlab Code Unif[0,1]

```
function[mu,variance,cov] = boot_iqrs2(dist,parm1,parm2,n,alpha)

tic;

r = floor((n+1)/4);

s = floor(3*(n+1)/4);

ratio1=mod(n+1,4)/4;

ratio2 = mod(3*(n+1),4)/4;

delta = icdf

(dist,.75,parm1,parm2) - icdf(dist,.25,parm1,parm2);

iqr_pop = [];

mn_boot = [];

var_boot = [];

lb_iqr=[];

ub_iqr=[];


for i = 1:2000


x=random(dist,parm1,parm2,n,1);

 x=sort(x);

 q1 = (x(r) + ratio1*(x(r+1)-x(r)));

 q3 = (x(s) + ratio2*(x(s+1)-x(s)));

 iqrange=q3-q1;

 u = random('unid',n,n,2000);
```

```
 boot_samples = x(u);
 sort_boot = sort(boot_samples);
 iqr_boot = (sort_boot(s,:)+ratio2*(sort_boot(s+1,:)-
 sort_boot(s,:))-sort_boot(r,:)+ratio1*(sort_boot(r+1,:)-sort_boot(r,:)));
 iqr_sort = sort(iqr_boot);
 lb = iqr_sort(2000*alpha/2);
 ub = iqr_sort(2000*(1-alpha/2) + 1);
 lb_iqr = [lb_iqr;lb];
 ub_iqr = [ub_iqr;ub];
 meaniqr = mean(iqr_boot);
 variqr = var(iqr_boot);
  iqr_pop = [iqr_pop;iqrange];
  mn_boot = [mn_boot; meaniqr];
  var_boot = [var_boot; variqr];



end
mu=mean(mn_boot);
variance=mean(var_boot);
cov = sum((lb_iqr
<= delta & delta <= ub_iqr))/2000;


toc;
```

VITA

Dewey L. Whaley III

628 Range Street

Elizabethton, TN 37643

EDUCATION

East Tennessee State University, Johnson City, TN

Mathematics, B.S., May 1999

East Tennessee State University, Johnson City, TN

Mathematics, M.S., August 2005

PROFESSIONAL EXPERIENCE

Math Lab Supervisor, Northeast State Technical Community College

Mathematics Department, August 2000 - August 2001

Teacher, Unaka High School

Mathematics Department, August 2001 - May 2004

Chief Financial Officer, Medical Care LLC

May 2004 - present

LICENSURE

Tennessee Teachers License, Mathematics "Highly Qualified"

Valid through 2014