SCHOOL of GRADUATE STUDIES
EAST TENNESSEE STATE UNIVERSITY

East Tennessee State University
Digital Commons @ East Tennessee State University

Electronic Theses and Dissertations | Student Works

8-2008

Estimating the Difference of Percentiles from Two Independent Populations.

Romual Eloge Tchouta
East Tennessee State University

Follow this and additional works at: https://dc.etsu.edu/etd

Part of the Statistical Theory Commons

Recommended Citation

Tchouta, Romual Eloge, "Estimating the Difference of Percentiles from Two Independent Populations." (2008). *Electronic Theses and Dissertations.* Paper 1981. https://dc.etsu.edu/etd/1981

This Thesis - Open Access is brought to you for free and open access by the Student Works at Digital Commons @ East Tennessee State University. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ East Tennessee State University. For more information, please contact digilib@etsu.edu.

Estimating the Difference of Percentiles from Two Independent Populations

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

Romual E. Tchouta

August 2008

Bob Price, Ph.D., Chair

Robert Gardner, Ph.D.

Yali Liu, Ph.D.

Keywords: confidence interval, percentile, normal distribution, exponential distribution, uniform distribution, empirical coverage.

ABSTRACT

Estimating the Difference of Percentiles from Two Independent Populations

by

Romual E. Tchouta

We first consider confidence intervals for a normal percentile, an exponential percentile and a uniform percentile. Then we develop confidence intervals for a difference of percentiles from two independent normal populations, two independent exponential populations and two independent uniform populations. In our study, we mainly focus on the maximum likelihood to develop our confidence intervals. The efficiency of this method is examined via coverage rates obtained in a simulation study done with the statistical software R.

Copyright by

Romual E. Tchouta 2008

All rights reserved

DEDICATION

To my grandmothers Rose Kapentengam and Anastasie Kuitchou that both passed away in 2006. I miss you guys so much and I hope God is watching over you guys. Love you guys...

ACKNOWLEDGMENTS

First off, I would like to thank God for the many blessings and without whom this never would have been achieved. My second words of acknowledgment go to all the members of the staff in the department of Mathematics at ETSU for their words of encouragement. My special thanks go to Dr. Janice Huang for her support as far as my admission to ETSU and her inspirational advice. I also wish to acknowledge the support of my committee members, namely, Dr. Bob Price, Dr. Bob Gardner and Dr. Yali Liu. They have all been helpful throughout my two years at ETSU.

Finally, I would like to acknowledge the support of my friends, colleagues and above all my family in Cameroon and all over the world: my mom (*Rémé Tendance*), my dad (*Donka*), my big brother (*Willy*), my big sister (*Mamiton*), my niece (*Orné*), my cousins (*Yann* and *Fany*), my aunts (*Maman Eli*, *Maman Bébé*, *Tata Esther*, *Tata Denise, Tata Marceline* and *Tata Regine*), the Batchadji family and my uncles (*Tonton Maniac*, *Papa Emma*, *Tonton Ilias*, *Tonton Isidore* and *Tonton Hubain*).

CONTENTS

LIST OF TABLES

1 INTRODUCTION

Percentiles are very important in both descriptive and inferential data analysis. They are used to describe key aspects of a distribution such as central tendency and spread. The most common percentiles are listed in the five number summary: the minimum, the 25^{th} percentile (called the first quartile), the 50^{th} percentile (called the median), the 75^{th} percentile (called the third quartile) and the maximum. The inter-quartile range (which is the difference between the third and first quartiles) is often used as a measure of spread of a distribution. Percentiles are used in several fields of study. For example, standardized tests like SAT, GRE, GMAT, etc. often report a student's performance using percentiles[3]. The median household income is commonly cited in economic statistics. In insurance, percentiles are used to set premiums.

A considerable amount of work has been done on statistical inference for percentiles. Methods, tests and confidence intervals have been developed for situations when the underlying distribution is unknown: these are called distribution-free methods. Some of these include order statistics; see, for example, *Gibbons and Chakraborti*[2]. Another popular method that plays a useful role in computing is called bootstrapping; see for example, *Efron and Tibshirani*[5].

The study of the difference in percentiles may be of interest when we want to compare two populations in terms of percentiles. A good example to illustrate the need to estimate the difference in percentiles would be comparing the typical student's performances (e.g. 70^{th} percentiles) between 2 groups. Usually, we consider the difference in means to compare two groups. There has been work comparing medians between two independent groups; see, for example, *Price and Bonett*[1]. In this thesis, we consider three distributions: the normal distribution, the exponential distribution and the uniform distribution. For each of these, we want to find confidence intervals for the difference in percentiles when the underlying distributions are independent. We focus on maximum likelihood to develop an approximate $(1-\alpha)100\%$ confidence interval for the difference of percentiles. The form of the interval will be $estimator \pm z_{\alpha/2} * standard\ error$.

1.1 Basic Definitions

A population parameter is a value used to represent a certain quantifiable characteristic of a population. As an example, the family of normal distributions has two parameters, the mean and the variance. Other examples of parameters are the standard deviation, the median, percentiles, and proportions.

An estimator is any quantity calculated from the sample data which is used to give information about a population parameter. For example, the usual estimator of the population mean is $\hat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ where n is the size of the sample $X_1, X_2, \ldots, X_n$ taken from the population.

An estimator $\hat{\theta}$ of a parameter θ is said to be unbiased if the expectation, $E(\hat{\theta})$, of $\hat{\theta}$ is equal to θ. Otherwise, it is biased. For example, the sample mean $\overline{X}$ is an unbiased estimator of the population mean μ.

An estimator $\hat{\theta}$ of a parameter θ is said to be asymptotically unbiased if $E(\hat{\theta}) \rightarrow \theta$

as $n \to \infty$, where n is the sample size.

For a given proportion ρ, a confidence interval for a population parameter is an interval that is calculated from a random sample of the underlying population such that, if the sampling was repeated numerous times and the confidence interval re calculated from each sample according to the same method, proportion ρ of the confidence intervals would contain the parameter. For example, the interval $[a, b]$ is a 95% confidence interval for the population mean μ if by repetition, in 95% of the cases, μ lies between a and b.

A $(100p)^{th}$ percentile is a value, k_p, such that at most $(100p)\%$ of the observations are less than this value and at most $100(1-p)\%$ are greater. That is, given a random variable X with p.d.f. $f(x)$ and c.d.f. $F(x)$, the $(100p)^{th}$ percentile is the number k_p such that $p = \int_{-\infty}^{k_p} f(x)dx = F(k_p)$. For instance, the 65^{th} percentile is the value below which 65% of the observations may be found.

Let X be a random variable which has a normal distribution with mean μ and standard deviation σ. Then the probability density function of X is given by [7]

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

Let Y be a random variable which has an exponential distribution with mean θ. Then the probability density function of Y is given by [6]

$$g(y) = \frac{1}{\theta} e^{-y/\theta}, \qquad 0 \leq y < \infty. \tag{2}$$

Let Z be a random variable which has a uniform distribution with interval of support $[a, b]$. Then the probability density function of Z is given by

$$h(z) = \frac{1}{b-a}, \qquad a \leq z \leq b. \tag{3}$$

1.2 Maximum Likelihood Estimator

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with p.m.f. or p.d.f. denoted by $f(x; \theta_1, \theta_2, \ldots, \theta_m)$. Suppose that $(\theta_1, \theta_2, \ldots, \theta_m)$ is restricted to a parameter space Ω. Then the joint p.m.f. or p.d.f. of $X_1, X_2, \ldots, X_n$, namely

$L(\theta_1, \theta_2, \ldots, \theta_m) = f(x_1; \theta_1, \theta_2, \ldots, \theta_m) f(x_2; \theta_1, \theta_2, \ldots, \theta_m) \cdots f(x_m; \theta_1, \theta_2, \ldots, \theta_m)$

where $(\theta_1, \theta_2, \ldots, \theta_m) \in \Omega$, when regarded as a function of $\theta_1, \theta_2, \ldots, \theta_m$, is called the likelihood function.

Say $[u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, \ldots, x_n)]$ is that $m-$tuple in Ω that maximizes $L(\theta_1, \theta_2, \ldots, \theta_m)$. Then

$$\hat{\theta}_1 = u_1(X_1, X_2, \ldots, X_n)$$

$$\hat{\theta}_2 = u_2(X_1, X_2, \ldots, X_n)$$

$$\vdots$$

$$\hat{\theta}_m = u_m(X_1, X_2, \ldots, X_n)$$

are maximum likelihood estimators of $\theta_1, \theta_2, \ldots, \theta_m$, respectively; and the corresponding observed values of these statistics, namely $u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n)$ $, \ldots,$ $u_m(x_1, x_2, \ldots, x_n)$, are called maximum likelihood estimates. In many practical cases, these estimators (and estimates) are unique.

For many applications there is just one unknown parameter. In these cases the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta). \tag{4}$$

As an illustration, let $X_1, X_2, \ldots, X_n$ be a random sample from the geometric distribution with p.m.f. $f(x; p) = (1-p)^{x-1}p$, where $x = 1, 2, 3, \cdots$. The likelihood function is given by

$$\begin{aligned} L(p) &= (1-p)^{x_1-1}p(1-p)^{x_2-1}p \cdots (1-p)^{x_n-1}p \\ &= p^n(1-p)^{\sum\limits_{i=1}^{n} x_i - n}, \qquad 0 \leq p \leq 1. \end{aligned} \tag{5}$$

The natural logarithm of $L(p)$ is

$$\ln L(p) = n \ln p + \left(\sum_{i=1}^{n} x_i - n\right) \ln(1-p), \qquad 0 < p < 1. \tag{6}$$

Thus restricting p to $0 < p < 1$ so as to be able to take the derivative, we have

$$\frac{d \ln L(p)}{dp} = \frac{n}{p} - \frac{\sum\limits_{i=1}^{n} x_i - n}{1-p} = 0.$$

Solving for p, we obtain

$$p = \frac{n}{\sum\limits_{i=1}^{n} x_i} = \frac{1}{\bar{x}} \tag{7}$$

and this solution provides a maximum. So the maximum likelihood estimator of p is

$$\hat{p} = \frac{n}{\sum\limits_{i=1}^{n} X_i} = \frac{1}{\overline{X}}. \tag{8}$$

2 CONFIDENCE INTERVAL FOR THE DIFFERENCE OF PERCENTILES FROM TWO NORMAL DISTRIBUTIONS

2.1 Confidence Interval of a Normal Distribution Percentile

Let X be a random variable which has a normal distribution with mean μ and variance σ^2. Then the p.d.f. of X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty. \tag{9}$$

Let k_p denote the $(100p)^{th}$ percentile of X. Then

$$k_p = \mu + Z_p\sigma \tag{10}$$

where Z_p denotes the $(100p)^{th}$ percentile of the standard normal distribution $N(0,1)$ [3]. Since μ and σ are unknown, we need to find estimators for those parameters.

Proposition 2.1 *Given a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution $N(\mu, \sigma^2)$, the maximum likelihood estimator of μ is the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.*

<u>Proof.</u>

The likelihood function is given by

$$\begin{aligned} L(\mu) &= \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \cdots \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}} \end{aligned} \tag{11}$$

The natural logarithm of $L(\mu)$ is

$$\ln L(\mu) = n \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{12}$$

Thus taking the derivative of $\ln L(\mu)$ with respect to μ, we have

$$\begin{aligned} \frac{d\ln L(\mu)}{d\mu} &= \frac{2}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\ &= \frac{\sum\limits_{i=1}^{n} x_i - n\mu}{\sigma^2} \quad = \ 0, \qquad \sigma \neq 0. \end{aligned} \tag{13}$$

Solving for μ, we obtain

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{14}$$

and this provides a maximum. So the maximum likelihood estimator for μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \qquad \blacksquare \tag{15}$$

Moreover, $\overline{X}$ is an unbiased estimator of μ since $E(\overline{X}) = \mu$.

Lemma 2.2 *Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a normal distribution $N(\mu, \sigma^2)$. Then the distribution of $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$, where $\chi^2(n-1)$ is a Chi-square distribution with n-1 degrees of freedom* [6].

Proposition 2.3 *Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a normal distribution $N(\mu, \sigma^2)$. Then the sample variance $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ is an unbiased estimator of σ^2.*

Proof.

By *Lemma* 2.2, the distribution of $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$. Therefore

$$\begin{aligned} E\left(\frac{(n-1)S^2}{\sigma^2}\right) &= E\left(\chi^2(n-1)\right) \\ \frac{n-1}{\sigma^2}E(S^2) &= n-1 \\ E(S^2) &= \sigma^2 \qquad \blacksquare \end{aligned} \tag{16}$$

Proposition 2.4 *cS is an unbiased estimator for σ, where $c = \sqrt{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})/\ \Gamma(\frac{n}{2})$.*

Proof.

We need to show that $E(cS) = \sigma$, i.e. $E(S) = \frac{\sigma}{c}$. We know from *Lemma* 2.2 that $\chi^2(n-1) \sim \frac{(n-1)S^2}{\sigma^2}$. So, $\sqrt{\chi^2(n-1)} \sim \frac{\sqrt{n-1}S}{\sigma}$.

Let's find the p.d.f. of $Y = \sqrt{\chi^2(n-1)}$. Suppose $f(x)$ and $g(y)$ are p.d.f.'s of $\chi^2(n-1)$ and $\sqrt{\chi^2(n-1)}$ respectively. Then

$$f(x) = \frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}}x^{\frac{n-1}{2}-1}\ e^{-\frac{x}{2}}, \qquad 0 \le x < \infty. \tag{17}$$

Thus, by the change-of-variables technique we have,

$$g(y) = \frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}}(y^2)^{\frac{n-1}{2}-1}\ e^{-\frac{(y^2)}{2}}2y, \qquad 0 \le y < \infty. \tag{18}$$

Now,

$$\begin{aligned} E(Y) &= \int_0^\infty y\frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}}(y^2)^{\frac{n-1}{2}-1}\ e^{-\frac{y^2}{2}}2ydy \\ &= \int_0^\infty (y^2)^{\frac{1}{2}}\frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}}(y^2)^{\frac{n-3}{2}}\ e^{-\frac{y^2}{2}}2ydy \\ &= \int_0^\infty \frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}}(y^2)^{\frac{n-2}{2}}\ e^{-\frac{y^2}{2}}2ydy \end{aligned}$$

Letting $t = y^2$, $dt = 2ydy$ and we obtain,

$$\begin{aligned} E(Y) &= \int_0^\infty \frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}} t^{\frac{n-2}{2}}\ e^{-\frac{t}{2}} dt \\ &= \int_0^\infty \frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}} t^{\frac{n}{2}-1}\ e^{-\frac{t}{2}} dt \\ &= \frac{1}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}} \int_0^\infty t^{\frac{n}{2}-1}\ e^{-\frac{t}{2}} dt \\ &= \frac{\Gamma(\frac{n}{2})2^{\frac{n}{2}}}{\Gamma(\frac{n-1}{2})2^{\frac{n-1}{2}}} \underbrace{\int_0^\infty \frac{t^{\frac{n}{2}-1}\ e^{-\frac{t}{2}}}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} dt}_{1}. \end{aligned}$$

Hence,

$$E(Y) = \frac{\Gamma(\frac{n}{2})\sqrt{2}}{\Gamma(\frac{n-1}{2})} \tag{19}$$

Since $Y = \sqrt{\chi^2(n-1)} \sim \frac{\sqrt{n-1}S}{\sigma}$, $E(Y) = \frac{\sqrt{n-1}}{\sigma} E(S)$ and therefore,

$$\begin{aligned} E(S) &= \frac{\sigma E(Y)}{\sqrt{n-1}} \\ &= \frac{\sigma}{\sqrt{n-1}} \frac{\Gamma(\frac{n}{2})\sqrt{2}}{\Gamma(\frac{n-1}{2})} \\ &= \frac{\sigma}{\sqrt{\frac{n-1}{2}}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \\ &= \frac{\sigma}{\frac{\sqrt{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}} \\ &= \frac{\sigma}{c}. \qquad \blacksquare \end{aligned} \tag{20}$$

Thus by *Proposition* 2.1 and *Proposition* 2.4, an unbiased estimator for k_p is

$$\hat{k_p} = \overline{X} + Z_p c S \tag{21}$$

Theorem 2.5 *Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a normal distribution $N(\mu, \sigma^2)$ where μ and σ^2 are unknown. Then a $(1-\alpha)100\%$ confidence* *interval for the* $(100p)^{th}$ *percentile,* k_p*, is*

$$(\overline{X} + Z_p cS) \pm z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{1 + nZ_p^2(c^2 - 1)} \tag{22}$$

where $c = \sqrt{\frac{n-1}{2}} \Gamma(\frac{n-1}{2}) / \ \Gamma(\frac{n}{2})$ *and* $P(Z > z_{\alpha/2}) = \alpha/2$*.*

Proof.

A $(1-\alpha)100\%$ confidence interval for k_p is $\hat{k}_p \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{k}_p)}}$ and by (21) $\hat{k}_p = \overline{X} + Z_p cS$. So, all we need to show is that $\widehat{Var(\hat{k}_p)} = \frac{S^2}{n}\left(1 + nZ_p^2(c^2-1)\right)$.

$$\begin{aligned}
Var(\hat{k}_p) &= Var(\overline{X} + Z_p cS) \\
&= Var(\overline{X}) + (cZ_p)^2 Var(S) \\
&= \frac{\sigma^2}{n} + c^2 Z_p^2 \left[E(S^2) - (E(S))^2\right] \\
&= \frac{\sigma^2}{n} + c^2 Z_p^2 \left[\sigma^2 - \frac{\sigma^2}{c^2}\right] \qquad \textit{by (16) and (20)} \\
&= \frac{\sigma^2}{n}\left(1 + nc^2 Z_p^2 (1 - \frac{1}{c^2})\right) \\
&= \frac{\sigma^2}{n}\left(1 + nZ_p^2(c^2-1)\right) \qquad (23)
\end{aligned}$$

Thus an estimator for $Var(\hat{k}_p)$ is

$$\widehat{Var(\hat{k}_p)} = \frac{S^2}{n}\left(1 + nZ_p^2(c^2-1)\right) \qquad \blacksquare \tag{24}$$

2.2 Confidence Interval of the Difference of Percentiles from Two Normal Percentiles

In this section, we consider two independent normal distributions $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$. The objective is to construct an approximate confidence interval for $k_p - k_p^{'}$

where k_p and $k_p^{'}$ are the $(100p)^{th}$ percentiles of $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$ respectively. We will use the results obtained in the previous section.

Theorem 2.6 *Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be 2 independent random samples of sizes n and m from the two normal distributions $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$. Let k_p and $k_p^{'}$ be the $(100p)^{th}$ percentiles of $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$, respectively. An approximate $(1-\alpha)100\%$ confidence interval for $k_p - k_p^{'}$ is*

$$\begin{aligned}&\big((\overline{X} + Z_p c_n S_x) - (\overline{Y} + Z_p c_m S_y)\big) \pm \\ &\qquad z_{\alpha/2}\sqrt{\tfrac{S_x^2}{n}\left(1 + nZ_p^2(c_n^2 - 1)\right) + \tfrac{S_y^2}{m}\left(1 + mZ_p^2(c_m^2 - 1)\right)}\end{aligned} \tag{25}$$

where Z_p denotes the $(100p)^{th}$ percentile of the standard normal distribution $N(0,1)$, $c_n = \sqrt{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})/\ \Gamma(\frac{n}{2})$ and $c_m = \sqrt{\frac{m-1}{2}}\Gamma(\frac{m-1}{2})/\ \Gamma(\frac{m}{2})$.

<u>Proof.</u>

A $(1-\alpha)100\%$ confidence interval of $k_p - k_p^{'}$ is

$$\begin{aligned} I &= \widehat{k_p - k_p^{'}} \pm z_{\alpha/2}\sqrt{\widehat{Var(\widehat{k_p - k_p^{'}})}} & (26)\\ &= \hat{k}_p - \hat{k}_p^{'} \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{k}_p - \hat{k}_p^{'})}} & (27)\\ &= \hat{k}_p - \hat{k}_p^{'} \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{k}_p)} + \widehat{Var(\hat{k}_p^{'})}} & (28)\end{aligned}$$

Now, using the same concept as in equations (21) and (24) from the previous section we have

$$
\begin{aligned}
\hat{k}_p &= \overline{X} + Z_p c_n S_x && (29)\\
\hat{k}_p' &= \overline{Y} + Z_p c_m S_y && (30)\\
\widehat{Var(\hat{k}_p)} &= \frac{S_x^2}{n}\left(1 + nZ_p^2(c_n^2 - 1)\right) && (31)\\
\widehat{Var(\hat{k}_p')} &= \frac{S_y^2}{m}\left(1 + mZ_p^2(c_m^2 - 1)\right) && (32)
\end{aligned}
$$

and the result follows. $\blacksquare$

2.3 Simulation Results

A simulation study was conducted to evaluate the coverage probabilities for the 90%, 95% and 99% confidence intervals for the difference in percentiles from two normal populations. We used the statistical software R to simulate the random data 100,000 times (the R code is shown in Appendix A)[8]. The parameters for the two normal distributions were fixed as follows: $\mu_1 = 10$, $\sigma_1 = 1$, $\mu_2 = 15$ and $\sigma_2 = 4$.

Table 1: Empirical coverage Rates of 90%, 95% and 99% Confidence Intervals for Difference in Percentiles from Two Normal Populations.

percentiles	n	m	90%	95%	99%
$p = 0.25$	10	10	0.8731	0.9219	0.9699
	50	10	0.8698	0.9190	0.9661
	50	50	0.8971	0.9443	0.9866
	200	100	0.8987	0.9481	0.9883
	500	500	0.8991	0.9487	0.9893
$p = 0.5$	10	10	0.8691	0.9214	0.9729
	50	10	0.8652	0.9286	0.9708
	50	50	0.8952	0.9455	0.9827
	200	100	0.8956	0.9480	0.9887
	500	500	0.9001	0.9495	0.9896
$p = 0.75$	10	10	0.8729	0.9226	0.9705
	50	10	0.8715	0.9282	0.9663
	50	50	0.8944	0.9449	0.9865
	200	100	0.8969	0.9476	0.9884
	500	500	0.8988	0.9495	0.9894
$p = 0.9$	10	10	0.8776	0.9233	0.9659
	50	10	0.8742	0.9273	0.9614
	50	50	0.8957	0.9443	0.9846
	200	100	0.8970	0.9467	0.9874
	500	500	0.9001	0.9508	0.9895

3 CONFIDENCE INTERVAL FOR THE DIFFERENCE OF PERCENTILES FROM TWO EXPONENTIAL DISTRIBUTIONS

3.1 Confidence Interval for an Exponential Distribution Percentile

Let X be a random variable which has an exponential distribution with mean θ and variance θ^2. Then the p.d.f. of X is given by

$$f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, \qquad 0 \leq x < \infty. \tag{33}$$

The $(100p)^{th}$ percentile of X is the number k_p such that $F(k_p) = p$. That is,

$$\begin{aligned} \int_0^{k_p} \frac{1}{\theta}e^{-\frac{x}{\theta}}dx &= p \\ 1 - e^{-\frac{k_p}{\theta}} &= p. \end{aligned}$$

Solving for k_p we obtain

$$k_p = -\theta \ln(1-p). \tag{34}$$

But θ being an unknown parameter, we need to estimate it.

Proposition 3.1 *Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from an exponential distribution with mean θ. The sample mean $\overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ is the MLE of θ.*

<u>Proof.</u>

The likelihood function is given by

$$\begin{aligned} L(\theta) &= L(\theta; x_1, x_2, \ldots, x_n) \\ &= \left(\frac{1}{\theta}e^{-x_1/\theta}\right)\left(\frac{1}{\theta}e^{-x_2/\theta}\right)\cdots\left(\frac{1}{\theta}e^{-x_n/\theta}\right) \\ &= \frac{1}{\theta^n}\exp\left(\frac{-\sum\limits_{i=1}^{n} x_i}{\theta}\right), \qquad 0 < \theta < \infty. \end{aligned} \tag{35}$$

The natural logarithm of $L(\theta)$ is

$$\ln L(\theta) = -n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} x_i, \qquad 0 < \theta < \infty. \tag{36}$$

Thus,

$$\frac{d\left[\ln L(\theta)\right]}{d\theta} = \frac{-n}{\theta} + \frac{\sum\limits_{i=1}^{n} x_i}{\theta^2} = 0$$

Solving for θ, we obtain

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{37}$$

Hence, the maximum likelihood estimator for θ is

$$\hat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \qquad \blacksquare \tag{38}$$

Also by the Central Limit Theorem, $\overline{X}$ is an unbiased estimator of θ. Thus an unbiased estimator for k_p is given by

$$\hat{k}_p = -\overline{X}\ln(1-p). \tag{39}$$

Theorem 3.2 *Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with unknown mean θ. Then a $(1-\alpha)100\%$ confidence interval for the $(100p)^{th}$ percentile, k_p, is given by*

$$-\overline{X}\ln(1-p) \pm z_{\alpha/2}|\ln(1-p)|\frac{\overline{X}}{\sqrt{n}}. \tag{40}$$

<u>Proof.</u>

A $(1-\alpha)100\%$ confidence interval for k_p is $\hat{k}_p \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{k}_p)}}$ and by (39) $\hat{k}_p = -\overline{X}\ln(1-p)$. Also,

$$\begin{aligned} Var(\hat{k}_p) &= Var(-\overline{X}\ln(1-p)) \\ &= (\ln(1-p))^2\, Var(\overline{X}) \\ &= (\ln(1-p))^2\, \frac{\theta^2}{n} \quad \text{by the Central Limit Theorem.} \end{aligned} \tag{41}$$

Thus,

$$\widehat{Var(\hat{k}_p)} = (\ln(1-p))^2\, \frac{\overline{X}^2}{n}. \tag{42}$$

Hence,

$$\sqrt{\widehat{Var(\hat{k}_p)}} = |\ln(1-p)|\frac{\overline{X}}{\sqrt{n}}. \tag{43}$$

Therefore a $(1-\alpha)100\%$ confidence interval for k_p is

$$-\overline{X}\ln(1-p) \pm z_{\alpha/2}|\ln(1-p)|\frac{\overline{X}}{\sqrt{n}}. \qquad \blacksquare \tag{44}$$

3.2 Confidence Interval for the Difference of Percentiles from Two Exponential Distributions

In this section, we will consider two exponential distributions D_1 and D_2 with respective unknown means θ_1 and θ_2. Our objective will be to find an approximate confidence interval of k_p - k'_p where k_p and k'_p denote the $(100p)^{th}$ percentiles of D_1 and D_2 respectively. For that purpose, we will use the results obtained on the previous section.

Theorem 3.3 *Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two independent random samples of sizes n and m from the two exponential distributions D_1 and D_2. Let k_p and k'_p be the $(100p)^{th}$ percentiles of D_1 and D_2 respectively. Then a $(1-\alpha)100\%$ confidence interval for $k_p - k'_p$ is given by*

$$\ln(1-p)\left(\overline{Y} - \overline{X}\right) \pm z_{\alpha/2} |\ln(1-p)| \sqrt{\frac{\overline{X}^2}{n} + \frac{\overline{Y}^2}{m}} \tag{45}$$

where $P(Z > z_{\alpha/2}) = \alpha/2$.

<u>Proof.</u>

By equation (25), a $(1-\alpha)100\%$ confidence interval for $k_p - k'_p$ is given by

$$\hat{k}_p - \hat{k}'_p \pm z_{\alpha/2} \sqrt{\widehat{Var(\hat{k}_p)} + \widehat{Var(\hat{k}'_p)}}. \tag{46}$$

From the results obtained in the previous section, we can establish the following equations :

$$\begin{aligned} \hat{k}_p &= -\overline{X} \ln(1-p) & (47) \\ \hat{k}'_p &= -\overline{Y} \ln(1-p) & (48) \\ \widehat{Var(\hat{k}_p)} &= (\ln(1-p))^2 \frac{\overline{X}^2}{n} & (49) \\ \widehat{Var(\hat{k}'_p)} &= (\ln(1-p))^2 \frac{\overline{Y}^2}{m}. & (50) \end{aligned}$$

Thus a $(1-\alpha)100\%$ confidence interval for $k_p - k'_p$ is

$$(-\overline{X} \ln(1-p)) - (-\overline{Y} \ln(1-p)) \pm z_{\alpha/2} \sqrt{(\ln(1-p))^2 \frac{\overline{X}^2}{n} + (\ln(1-p))^2 \frac{\overline{Y}^2}{m}}.$$

And factoring out $\ln(1-p)$ and $|\ln(1-p)|$, we obtain

$$\ln(1-p)\left(\overline{Y} - \overline{X}\right) \pm z_{\alpha/2} |\ln(1-p)| \sqrt{\frac{\overline{X}^2}{n} + \frac{\overline{Y}^2}{m}}. \qquad \blacksquare$$

3.3 Simulation Results

A simulation study was conducted to evaluate the coverage probabilities for the 90%, 95% and 99% confidence intervals for the difference in percentiles from two exponential populations. We used the statistical software R to simulate the random data 100,000 times (the R code is shown in Appendix B)[8]. We fixed the parameters of the exponential distributions to be $\theta_1 = 10$ and $\theta_2 = 15$.

Table 2: Empirical Coverage Rates of 90%, 95% and 99% Confidence Intervals for Difference in Percentiles from Two Exponential populations.

percentiles	n	m	90%	95%	99%
	10	10	0.9181	0.9632	0.9931
	50	10	0.8779	0.9167	0.9597
$p = 0.25$	50	50	0.9051	0.9534	0.9908
	200	100	0.8999	0.9479	0.9887
	500	500	0.9003	0.9503	0.9898
	10	10	0.9163	0.9626	0.9933
	50	10	0.8803	0.9168	0.9591
$p = 0.5$	50	50	0.9033	0.9544	0.9910
	200	100	0.8998	0.9500	0.9875
	500	500	0.8997	0.9511	0.9901
	10	10	0.9181	0.9621	0.9934
	50	10	0.8815	0.9183	0.9595
$p = 0.75$	50	50	0.9030	0.9538	0.9907
	200	100	0.8987	0.9497	0.9884
	500	500	0.9017	0.9502	0.9907
	10	10	0.9178	0.9619	0.9931
	50	10	0.8796	0.9176	0.9587
$p = 0.9$	50	50	0.9038	0.9540	0.9910
	200	100	0.9038	0.9540	0.9877
	500	500	0.9002	0.9513	0.9901

4 CONFIDENCE INTERVAL FOR THE DIFFERENCE OF PERCENTILES FROM TWO UNIFORM DISTRIBUTIONS

4.1 Confidence Interval for a Uniform Distribution Percentile

Let X be a random variable which has a uniform distribution with interval of support $[a, b]$. Then the p.d.f. of X is given by

$$f(x) = \frac{1}{b-a}, \qquad a \leq X \leq b. \tag{51}$$

The $(100p)^{th}$ percentile of X is the number k_p such that is $F(k_p) = p$. That is,

$$\begin{aligned} \int_a^{k_p} \frac{1}{b-a} &= p \\ \frac{k_p - a}{b-a} &= p. \end{aligned}$$

Solving for k_p, we have

$$k_p = a + p(b-a). \tag{52}$$

Thus, an estimator for k_p is given by

$$\hat{k}_p = \hat{a} + p(\hat{b} - \hat{a}). \tag{53}$$

We will use $X_{(1)}$ and $X_{(n)}$ as estimators for a and b. So let's establish the following proposition.

Proposition 4.1 *Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution $U(a, b)$ where $[a, b]$ is the interval of support. Let the random variables $X_{(1)}, X_{(2)}$*

$, \ldots, X_{(n)}$ denote the order statistics of that sample. That is,

$$\begin{aligned} X_{(1)} &= \textit{smallest of } X_1, X_2, \ldots, X_n \\ X_{(2)} &= \textit{second smallest of } X_1, X_2, \ldots, X_n \\ &\vdots \\ X_{(n)} &= \textit{largest of } \ X_1, X_2, \ldots, X_n. \end{aligned}$$

Then,

i) $\hat{a} = X_{(1)}$ is an asymptotically unbiased estimator for a

ii) $\hat{b} = X_{(n)}$ is an asymptotically unbiased estimator for b.

<u>Proof.</u>

We need to show that $\lim_{n\to\infty} E(X_{(1)}) = a$ and $\lim_{n\to\infty} E(X_{(n)}) = b$.

i) We first show that $\lim_{n\to\infty} E(X_{(1)}) = a$. But before doing that, note that the p.d.f. of $X_{(1)}$ is given by

$$g_1(y) = n\left[1 - F(y)\right]^{n-1} f(y), \qquad a < y < b \tag{54}$$

where $f(y)$ is the p.d.f. of the X_i's and $F(y)$ is the c.d.f. of the X_i's. In this case we have

$$f(y) = \frac{1}{b-a}, \qquad a < y < b \tag{55}$$

and

$$\begin{aligned} F(y) &= \int_a^y \frac{1}{b-a} dy \\ &= \frac{y-a}{b-a}. \end{aligned} \tag{56}$$

Thus,

$$\begin{aligned} g_1(y) &= n\left(1-\frac{y-a}{b-a}\right)^{n-1}\frac{1}{b-a} \\ &= n\left(\frac{b-y}{b-a}\right)^{n-1}\frac{1}{b-a} \\ &= \frac{n(b-y)^{n-1}}{(b-a)^n}. \end{aligned} \tag{57}$$

Now, the expectation of $X_{(1)}$ is given by

$$\begin{aligned} E(X_{(1)}) &= \int_a^b y g_1(y) dy \\ &= \int_a^b y\frac{n(b-y)^{n-1}}{(b-a)^n}dy \\ &= \frac{n}{(b-a)^n}\int_a^b y(b-y)^{n-1}dy. \end{aligned}$$

Using integration by parts with $u = y$, $du = 1$ and $dv = (b-y)^{n-1}$, $v = \frac{-(b-y)^n}{n}$, we obtain

$$\begin{aligned} E(X_{(1)}) &= \frac{n}{(b-a)^n}\left[\frac{-y(b-y)^n}{n} - \frac{(b-y)^{n+1}}{n(n+1)}\right]_a^b \\ &= \frac{1}{(b-a)^n}\left[-y(b-y)^n - \frac{(b-y)^{n+1}}{n+1}\right]_a^b \\ &= \frac{1}{(b-a)^n}\left(0-\left(-a(b-a)^n - \frac{(b-a)^{n+1}}{n+1}\right)\right) \\ &= \frac{a(b-a)^n}{(b-a)^n} + \frac{(b-a)^{n+1}}{(n+1)(b-a)^n} \\ &= a + \frac{b-a}{n+1}. \end{aligned}$$

Hence,

$$\lim_{n\to\infty} E(X_{(1)}) = \lim_{n\to\infty} a + \frac{b-a}{n+1} = a. \tag{58}$$

ii) We now show that $\lim_{n\to\infty} E(X_{(n)}) = b$. Note that the p.d.f. of $X_{(n)}$ is given by

$$\begin{aligned} g_n(y) &= n\,[F(y)]^{n-1} f(y), \qquad a < y < b & (59)\\ &= n\left(\frac{y-a}{b-a}\right)^{n-1} \frac{1}{b-a} \qquad \textit{by } (52) \textit{ and } (53) \\ &= \frac{n(y-a)^{n-1}}{(b-a)^n}. & (60)\end{aligned}$$

Thus, the expectation of $X_{(n)}$ is given by

$$\begin{aligned} E(X_{(n)}) &= \int_a^b y g_n(y) dy \\ &= \int_a^b y \frac{n(y-a)^{n-1}}{(b-a)^n} dy \\ &= \frac{n}{(b-a)^n} \int_a^b y(y-a)^{n-1} dy. \end{aligned}$$

Using integration by parts with $u = y$, $du = 1$, $dv = (y-a)^{n-1}$, $v = \frac{(y-a)^n}{n}$, we have

$$\begin{aligned} E(X_{(n)}) &= \frac{n}{(b-a)^n}\left[\frac{y(y-a)^n}{n} - \frac{(y-a)^{n+1}}{n(n+1)}\right]_a^b \\ &= \frac{1}{(b-a)^n}\left[y(y-a)^n - \frac{(y-a)^{n+1}}{n+1}\right]_a^b \\ &= \frac{1}{(b-a)^n}\left(b(b-a)^n - \frac{(b-a)^{n+1}}{n+1} - 0\right) \\ &= \frac{b(b-a)^n}{(b-a)^n} - \frac{(b-a)^{n+1}}{(n+1)(b-a)^n} \\ &= b - \frac{b-a}{n+1}. \end{aligned}$$

Hence,

$$\lim_{n\to\infty} E(X_{(n)}) = \lim_{n\to\infty} b - \frac{b-a}{n+1} = b. \qquad \blacksquare \tag{61}$$

Therefore an estimator for k_p is given by

$$\hat{k}_p = X_{(1)} + (X_{(n)} - X_{(1)})p. \tag{62}$$

Theorem 4.2 *Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution $U(a,b)$ where $[a,b]$ is the interval of support. Let the random variables $X_{(1)}, X_{(2)}, \ldots,$ $X_{(n)}$ denote the order statistics of that sample. That is,*

$$\begin{aligned} X_{(1)} &= \textit{smallest of } X_1, X_2, \ldots, X_n \\ X_{(2)} &= \textit{second smallest of } X_1, X_2, \ldots, X_n \\ &\vdots \\ X_{(n)} &= \textit{largest of } \ X_1, X_2, \ldots, X_n. \end{aligned}$$

Then a $(1-\alpha)100\%$ confidence interval of the $(100p)^{th}$ percentile, k_p, of $U(a,b)$ is given by

$$X_{(1)} + \left(X_{(n)} - X_{(1)}\right) p \pm z_{\alpha/2} \frac{X_{(n)} - X_{(1)}}{n+1} \sqrt{\frac{2p^2(n-1) - 2p(n-1) + n}{n+2}}. \tag{63}$$

Proof.

A $(1-\alpha)$ confidence interval for k_p is $\hat{k}_p \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{k}_p)}}$. By equation (59), $\hat{k}_p = X_{(1)} + (X_{(n)} - X_{(1)})p$; thus all we need to show is that

$$\widehat{Var(\hat{k}_p)} = \frac{(X_{(n)} - X_{(1)})^2}{(n+1)^2(n+2)} \left(2p^2(n-1) - 2p(n-1) + n\right). \tag{64}$$

Now,

$$\begin{aligned} Var(\hat{k}_p) &= Var\left(X_{(1)} + (X_{(n)} - X_{(1)})p\right) \\ &= Var\left(X_{(1)} + X_{(n)}p - X_{(1)}p\right) \\ &= Var\left((1-p)X_{(1)} + pX_{(n)}\right) \\ &= (1-p)^2 Var\left(X_{(1)}\right) + p^2 Var\left(X_{(n)}\right) + 2p(1-p)Cov\left(X_{(1)}, X_{(n)}\right). \end{aligned} \tag{65}$$

Let's find $Var\left(X_{(1)}\right)$, $Var\left(X_{(n)}\right)$ and $Cov\left(X_{(1)}, X_{(n)}\right)$.

$$\begin{aligned} Var(X_{(1)}) &= E(X_{(1)}^2) - (E(X_{(1)}))^2 \\ &= \int_a^b y^2 \frac{n(b-y)^{n-1}}{(b-a)^n} dy - \left(a + \frac{b-a}{n+1}\right)^2. \end{aligned} \tag{66}$$

Using integration by parts we have

$$\int_a^b y^2 \frac{n(b-y)^{n-1}}{(b-a)^n} dy = a^2 + \frac{2a(b-a)}{n+1} + \frac{2(b-a)^2}{(n+1)(n+2)}. \tag{67}$$

So plugging equation (67) into equation (66) we obtain

$$\begin{aligned} Var(X_{(1)}) &= a^2 + \frac{2a(b-a)}{n+1} + \frac{2(b-a)^2}{(n+1)(n+2)} - \left(a^2 + \frac{2a(b-a)}{n+1} + \frac{(b-a)^2}{(n+1)^2}\right) \\ &= \frac{2(b-a)^2}{(n+1)(n+2)} - \frac{(b-a)^2}{(n+1)^2} \\ &= \frac{n(b-a)^2}{(n+1)^2(n+2)}. \end{aligned} \tag{68}$$

Also,

$$\begin{aligned} Var(X_{(n)}) &= E(X_{(n)}^2) - (E(X_{(n)}))^2 \\ &= \int_a^b \frac{ny^2(y-a)^{n-1}}{(b-a)^n} dy - \left(b - \frac{b-a}{n+1}\right)^2. \end{aligned} \tag{69}$$

Using integration by parts, we obtain

$$\int_a^b \frac{ny^2(y-a)^{n-1}}{(b-a)^n} dy = b^2 - \frac{2b(b-a)}{n+1} + \frac{2(b-a)^2}{(n+1)(n+2)}. \tag{70}$$

Plugging equation (70) into (69) we have

$$\begin{aligned} Var(X_{(n)}) &= b^2 - \frac{2b(b-a)}{n+1} + \frac{2(b-a)^2}{(n+1)(n+2)} - \left(b^2 - \frac{2b(b-a)}{n+1} + \frac{(b-a)^2}{(n+1)^2}\right) \\ &= \frac{2(b-a)^2}{(n+1)(n+2)} - \frac{(b-a)^2}{(n+1)^2} \\ &= \frac{n(b-a)^2}{(n+1)^2(n+2)}. \end{aligned} \tag{71}$$

So we observe that $Var(X_{(1)}) = Var(X_{(n)}) = \frac{n(b-a)^2}{(n+1)^2(n+2)}$. To find $Cov\left(X_{(1)}, X_{(n)}\right)$, we need to evaluate the joint probability of $X_{(1)}$ and $X_{(n)}$, $f_{X_{(1)},X_{(n)}}$, since we know that

$$\begin{aligned} Cov(X_{(1)}, X_{(n)}) &= E(X_{(1)}, X_{(n)}) - E(X_{(1)})E(X_{(n)}) \\ &= \int_a^b \int_a^y xy f_{X_{(1)},X_{(n)}} dxdy - E(X_{(1)})E(X_{(n)}). \end{aligned} \tag{72}$$

Consider a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with interval of support $[a, b]$ which has p.d.f. $f(x)$ and c.d.f. $F(x)$. Let the random variables $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of that sample. Then, the joint distribution of any 2 order statistics $X_{(i)}$ and $X_{(j)}$ is given by [4]

$$\begin{aligned} f_{X_{(i)},X_{(j)}} = &\frac{n!}{(i-1)!(j-i-1)!(n-j)!} \left(F(x)\right)^{i-1} f(x) \\ &\times \left(F(x) - F(y)\right)^{j-i-1} f(y) \left(1 - F(y)\right)^{n-j}. \end{aligned} \tag{73}$$

For $i = 1$ and $j = n$, we have

$$\begin{aligned} f_{X_{(1)},X_{(n)}} &= \frac{n!}{(1-1)!(n-1-1)!(n-n)!} \left(F(x)\right)^{1-1} f(x) \\ &\qquad \times \left(F(x) - F(y)\right)^{n-1-1} f(y) \left(1 - F(y)\right)^{n-n} \\ &= \frac{n!}{0!(n-2)!0!} \left(F(x)\right)^0 f(x) \left(F(x) - F(y)\right)^{n-2} f(y) \left(1 - F(y)\right)^0 \\ &= \frac{n!}{(n-2)!} \left(\frac{x-a}{b-a} - \frac{y-a}{b-a}\right)^{n-2} \\ &= n(n-1)\frac{(x-y)^{n-2}}{(b-a)^n}. \end{aligned} \tag{74}$$

Thus,

$$\begin{aligned}
Cov(X_{(1)}, X_{(n)}) &= \int_a^b \int_a^y xyn(n-1)\frac{(x-y)^{n-2}}{(b-a)^n}dxdy - E(X_{(1)})E(X_{(n)}) \\
&= \frac{n(n-1)}{(b-a)^n}\int_a^b \int_a^y xy(x-y)^{n-2}dxdy - E(X_{(1)})E(X_{(n)}).
\end{aligned}$$

Evaluating the double integral above using integration by parts, we have

$$\begin{aligned}
\int_a^b \int_a^y xy(x-y)^{n-2}dxdy = & \tfrac{ab(b-a)^n}{n(n-1)} - \tfrac{a(b-a)^{n+1}}{n(n-1)(n+1)} - \tfrac{b(b-a)^{n+1}}{n(n-1)(n+1)} \\
& - \tfrac{(b-a)^{n+2}}{n(n-1)(n+1)(n+2)}.
\end{aligned}$$

Thus,

$$\frac{n(n-1)}{(b-a)^n}\int_a^b \int_a^y xy(x-y)^{n-2}dxdy = ab + \frac{(b-a)^2}{n+2}. \tag{75}$$

Also,

$$\begin{aligned}
E(X_{(1)})E(X_{(n)}) &= \left(a + \frac{b-a}{n+1}\right)\left(b - \frac{b-a}{n+1}\right) \\
&= ab - \frac{a(b-a)}{n+1} + \frac{b(b-a)}{n+1} - \frac{(b-a)^2}{(n+1)^2} \\
&= ab + \frac{(b-a)^2}{n+1} - \frac{(b-a)^2}{(n+1)^2} \\
&= ab + \frac{n(b-a)^2}{(n+1)^2}. \qquad (76)
\end{aligned}$$

Combining equations (75) and (76), we have

$$\begin{aligned}
Cov(X_{(1)}, X_{(n)}) &= ab + \frac{(b-a)^2}{n+2} - ab - \frac{n(b-a)^2}{(n+1)^2} \\
&= \frac{(b-a)^2}{n+2} - \frac{n(b-a)^2}{(n+1)^2} \\
&= \frac{(b-a)^2}{(n+1)^2(n+2)}(n+1-n) \\
&= \frac{(b-a)^2}{(n+1)^2(n+2)}. \qquad (77)
\end{aligned}$$

Hence, plugging equations (68), (71) and (77) into (65) we obtain

$$
\begin{aligned}
Var(\hat{k}_p) &= (1-p)^2\frac{n(b-a)^2}{(n+1)^2(n+2)} + p^2\frac{n(b-a)^2}{(n+1)^2(n+2)} + 2p(1-p)\frac{(b-a)^2}{(n+1)^2(n+2)} \\
&= \frac{(b-a)^2}{(n+1)^2(n+2)}\left(n(1-p)^2 + np^2 + 2p(1-p)\right) \\
&= \frac{(b-a)^2}{(n+1)^2(n+2)}\left(n - 2np + np^2 + np^2 + 2p - 2p^2\right) \\
&= \frac{(b-a)^2}{(n+1)^2(n+2)}\left(2p^2(n-1) - 2p(n-1) + n\right). \qquad (78)
\end{aligned}
$$

Therefore,

$$
\widehat{Var(\hat{k}_p)} = \frac{(X_{(n)} - X_{(1)})^2}{(n+1)^2(n+2)}\left(2p^2(n-1) - 2p(n-1) + n\right). \qquad (79)
$$

and the result follows. ■

4.2 Confidence Interval of the Difference of Percentiles from Two Uniform Distributions

In this section, we consider two uniform distributions $U(a,b)$ and $U(c,d)$ with intervals of support $[a,b]$ and $[c,d]$, respectively. The objective is to find an approximate confidence interval for $k_p - k'_p$ where k_p and k'_p are the $(100p)^{th}$ percentiles of $U(a,b)$ and $U(c,d)$, respectively. We will apply the results from the previous section to construct an approximate confidence interval for the difference of percentiles.

Theorem 4.3 *Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two independent random samples of sizes n and m from two uniform distributions U(a,b) and U(c,d) with intervals of support $[a,b]$ and $[c,d]$ respectively. Let the random variables $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ and $Y_{(1)}, Y_{(2)}, \ldots, Y_{(m)}$ denote the order statistics of the first and second samples respectively. Let k_p and k'_p be the $(100p)^{th}$ percentiles of $U(a,b)$ and $U(c,d)$ respectively.*

Then a $(1-\alpha)100\%$ confidence interval for $k_p - k'_p$ is given by

$$\begin{aligned}
&(X_{(1)} - Y_{(1)}) + \left[\left(X_{(n)} - Y_{(m)}\right) - \left(X_{(1)} - Y_{(1)}\right)\right] p \pm z_{\alpha/2} \\
&\times\Big(\tfrac{(X_{(n)}-X_{(1)})^2}{(n+1)^2(n+2)} \left(2p^2(n-1) - 2p(n-1) + n\right) + \\
&\qquad \tfrac{(Y_{(n)}-Y_{(1)})^2}{(m+1)^2(m+2)} \left(2p^2(m-1) - 2p(m-1) + m\right) \Big)^{1/2} \qquad (80)
\end{aligned}$$

where $P(Z > z_{\alpha/2}) = \alpha/2$.

Proof.

By equation (28), a $(1-\alpha)100\%$ confidence interval for $k_p - k'_p$ is given by

$$\hat{k}_p - \hat{k}'_p \pm z_{\alpha/2}\sqrt{\widehat{Var(\hat{k}_p)} + \widehat{Var(\hat{k}'_p)}}. \qquad (81)$$

The two distributions in question being independent, we have $Cov(X_{(i)}, Y_{(j)}) = 0$ for any $i = 1, \ldots, n$ and $j = 1, \ldots, m$. In particular,

$$\begin{aligned}
Cov(X_{(1)}, Y_{(1)}) &= 0 \qquad (82) \\
Cov(X_{(1)}, Y_{(m)}) &= 0 \qquad (83) \\
Cov(X_{(n)}, Y_{(1)}) &= 0 \qquad (84) \\
Cov(X_{(n)}, Y_{(m)}) &= 0. \qquad (85)
\end{aligned}$$

Therefore, we can use equation (81) to get our desired confidence interval. From the results obtained in the previous section, we can establish the following equations :

$$\begin{aligned}
\hat{k}_p &= X_{(1)} + \left(X_{(n)} - X_{(1)}\right) p \qquad (86) \\
\hat{k}'_p &= Y_{(1)} + \left(Y_{(m)} - Y_{(1)}\right) p \qquad (87) \\
\widehat{Var(\hat{k}_p)} &= \frac{(X_{(n)} - X_{(1)})^2}{(n+1)^2(n+2)} \left(2p^2(n-1) - 2p(n-1) + n\right) \qquad (88) \\
\widehat{Var(\hat{k}'_p)} &= \frac{(Y_{(n)} - Y_{(1)})^2}{(m+1)^2(m+2)} \left(2p^2(m-1) - 2p(m-1) + m\right). \qquad (89)
\end{aligned}$$

So,

$$\begin{aligned}
\hat{k}_p - \hat{k}'_p &= \left[X_{(1)} + (X_{(n)} - X_{(1)})p\right] - \left[Y_{(1)} + (Y_{(m)} - Y_{(1)})p\right] \\
&= \left(X_{(1)} - Y_{(1)}\right) + \left[\left(X_{(n)} - X_{(1)}\right)p - \left(Y_{(m)} - Y_{(1)}\right)p\right] \\
&= \left(X_{(1)} - Y_{(1)}\right) + \left[\left(X_{(n)} - Y_{(m)}\right) - \left(X_{(1)} - Y_{(1)}\right)\right]p, \qquad (90)
\end{aligned}$$

and, adding equations (88) and (89) we have $\widehat{Var(\hat{k}_p)} + \widehat{Var(\hat{k}'_p)} =$

$$\begin{aligned}
&\tfrac{(X_{(n)}-X_{(1)})^2}{(n+1)^2(n+2)}\left(2p^2(n-1) - 2p(n-1) + n\right) + \\
&\qquad \tfrac{(Y_{(n)}-Y_{(1)})^2}{(m+1)^2(m+2)}\left(2p^2(m-1) - 2p(m-1) + m\right). \qquad \blacksquare
\end{aligned}$$

4.3 Simulation Results

The simulation study was conducted to estimate the coverage rates for the 90%, 95% and 99% confidence intervals for the difference in percentiles from two normal populations. We used the statistical software R to generate the random data and simulate the values 100,000 times (the R code is shown in Appendix C)[8]. The intervals of support of the distributions were fixed as follows: $[a, b]$=$[2, 4]$ and $[c, d]$=$[3, 5]$.

Table 3: Empirical Coverage Rates of 90%, 95% and 99% Confidence Intervals for the Difference in Percentiles from Two Uniform Populations

percentiles	n	m	90%	95%	99%
$p = 0.25$	10	10	0.8228	0.8752	0.9369
	50	10	0.7996	0.8437	0.9014
	50	50	0.8893	0.9282	0.9695
	200	100	0.8917	0.9282	0.9661
	500	500	0.9001	0.9367	0.9741
$p = 0.5$	10	10	0.8183	0.8732	0.9385
	50	10	0.8131	0.8614	0.9204
	50	50	0.8862	0.9289	0.9728
	200	100	0.8945	0.9340	0.9733
	500	500	0.9006	0.9403	0.9779
$p = 0.75$	10	10	0.8243	0.8735	0.9367
	50	10	0.8017	0.8456	0.9010
	50	50	0.8908	0.9276	0.9698
	200	100	0.8928	0.9265	0.9658
	500	500	0.9004	0.9380	0.9751
$p = 0.9$	10	10	0.8259	0.8756	0.9348
	50	10	0.7725	0.8177	0.8792
	50	50	0.8882	0.9268	0.9681
	200	100	0.8847	0.9202	0.9598
	500	500	0.9001	0.9364	0.9729

5 CONCLUSION

We observe from Table 1, Table 2 and Table 3 that with small values of n and m (for example $n = 10$ and $m = 10$ or $n = 50$ and $m = 10$), the coverage probabilities can be on the liberal side. However, as both n and m increase, the coverage probabilities converge to the desired nominal level. It is important to note that, in this thesis, the underlying distributions were known in advance. A possible alternative method for estimating the difference between percentiles from two independent groups when the underlying distributions are unknown would be bootstrapping which is a computer intensive method based on resampling. This could be considered as a direction for future research.

BIBLIOGRAPHY

[1] R. M. Price and D. G. Bonett, *J. Statist. Comput. Simul.*. Taylor & Francis Ltd, 2002, Vol. 72(2), 119-124.

[2] J. D. Gibbons and S. Chakraborti, *Non Parametric Statistical inference*, 4th edition. Marcel Dekker, New York, 2003.

[3] S. Charakraborti and J. Li, *Confidence Interval Estimation of a Normal Percentile*. American Statistical Association November 2007,Vol.61, No.4, 331-336.

[4] Herbert A. David, *Order Statistics*, 2nd edition. A Wiley Publication in applied statistics, 1981, 10-11.

[5] B. Efron and R. J, Tibshirani, *Introduction to the Bootstrap*. Chapman and Hall, 1993.

[6] R. V. Hogg and E. A. Tanis, *Probability and Statistical Inference*, 7th edition. Pearson Education, Inc. 2006.

[7] W. Mendenhall and R. L. Scheaffer, *Mathematical Statistics with Applications*. Wadsworth Publishing Company, Inc., 1973, 124-125.

[8] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Austria, Vienna, 2008.

APPENDICES

1 Appendix A: R Code for the Empirical Coverage Rates of Confidence Intervals for the Difference in Percentiles from Two Normal Distributions.

```
norm_cover = function(nsims,n,m,mu1,sig1,mu2,sig2,alpha,p) {
zp = qnorm(p)
kp1 = mu1 + zp*sig1
kp2 = mu2 + zp*sig2
diff = kp1 - kp2
cc = 1 - alpha
ic = 0
for ( i in 1:nsims) {
samp1 = rnorm(n,mu1,sig1)
samp2 = rnorm(m,mu2,sig2)
#c1 = (sqrt((n-1)/2) * gamma((n-1)/2)) / (gamma(n/2))     When n -> inf
gamma function fails; use log gamma
lnc1 = log(sqrt((n-1)/2)) + lgamma((n-1)/2) - lgamma(n/2)
c1 = exp(lnc1)
#c2 = (sqrt((m-1)/2) * gamma((m-1)/2)) / (gamma(m/2))
lnc2 = log(sqrt((m-1)/2)) + lgamma((m-1)/2) - lgamma(m/2)
c2 = exp(lnc2)
mean1 = mean(samp1)
sd1 = sd(samp1)
```

```
mean2 = mean(samp2)
sd2 = sd(samp2)
kp1_hat = mean1 + c1*zp*sd1
kp2_hat = mean2 + c2*zp*sd2
var_kp1_hat = (sd1^ 2/n)*(1 + n*zp^ 2*(c1^ 2 - 1) # estimated variance of
kp1_hat
var_kp2_hat = (sd2^ 2/m)*(1 + m*zp^ 2*(c2^ 2 - 1)) # estimated variance
of kp2_hat
crit = qnorm(1-alpha/2)
lb = kp1_hat - kp2_hat - crit*sqrt(var_kp1_hat + var_kp2_hat)
ub = kp1_hat - kp2_hat + crit*sqrt(var_kp1_hat + var_kp2_hat)
if (lb <= diff  &  diff <= ub)   {ic = ic + 1}
}
empcov = ic/nsims
list( empiricalcover = empcov )
}
```

2 Appendix B: R Code for the Empirical Coverage Rates of Confidence Intervals for the Difference in Percentiles from Two Exponential Distributions.

```
expo_cover = function(nsims,n,m,theta1,theta2,alpha,p){
kp1 = qexp(p,theta1)
kp2 = qexp(p,theta2)
diff = kp1 - kp2
cc = 1- alpha
ic = 0
for (i in 1:nsims) {
samp1 = rexp(n,theta1)
samp2 = rexp(m,theta2)
mean1 = mean(samp1)
mean2 = mean(samp2)
kp1_hat = - mean1 * log(1-p)
kp2_hat = - mean2 * log(1-p)
var_kp1_hat = (log(1-p))^ 2 * mean1^ 2 / n
var_kp2_hat = (log(1-p))^ 2 * mean2^ 2 / m
crit = qnorm(1-alpha/2)
lb = kp1_hat - kp2_hat - crit*sqrt(var_kp1_hat + var_kp2_hat)
ub = kp1_hat - kp2_hat + crit*sqrt(var_kp1_hat + var_kp2_hat)
if (lb <= diff & diff <= ub) {ic = ic + 1}
}
empcov = ic/nsims
```

```
list( empiricalcover = empcov )

}
```

3 Appendix C: R Code for the Empirical Coverage Rates of Confidence Intervals for the Difference in Percentiles from Two Uniform Distributions.

```
unif_cover = function(nsims,n,m,a,b,c,d,alpha,p) {
kp1 = a + (b-a)*p
kp2 = c + (d-c)*p
diff = kp1 - kp2
cc = 1 - alpha
ic = 0
for ( i in 1:nsims) {
samp1 = runif(n,a,b)
samp2 = runif(m,c,d)
ordered_samp1 = sort(samp1)
ordered_samp2 = sort(samp2)
a_hat = ordered_samp1[1]
b_hat = ordered_samp1[n]
c_hat = ordered_samp2[1]
d_hat = ordered_samp2[m]
kp1_hat = a_hat + (b_hat - a_hat)*p
kp2_hat = c_hat + (d_hat - c_hat)*p
var_a_hat = (n * (b_hat-a_hat)^ 2)/((n+2)*(n+1)^ 2)
var_b_hat = (n * (b_hat-a_hat)^ 2)/((n+2)*(n+1)^ 2)
var_c_hat = (m * (d_hat-c_hat)^ 2)/((m+2)*(m+1)^ 2)
var_d_hat = (m * (d_hat-c_hat)^ 2)/((m+2)*(m+1)^ 2)
```

```
cov1 = (b_hat - a_hat)^ 2 / ((n+2)*(n+1)^ 2)

cov2 = (d_hat - c_hat)^ 2 / ((m+2)*(m+1)^ 2)

var_kp1_hat = (1-p)^ 2 * var_a_hat + p^ 2 * var_b_hat + 2*p*(1-p)*cov1

var_kp2_hat = (1-p)^ 2 * var_c_hat + p^ 2 * var_d_hat + 2*p*(1-p)*cov2

crit = qnorm(1-alpha/2)

lb = kp1_hat - kp2_hat - crit*sqrt(var_kp1_hat + var_kp2_hat)

ub = kp1_hat - kp2_hat + crit*sqrt(var_kp1_hat + var_kp2_hat)

if (lb <= diff & diff <= ub) {ic = ic + 1}

}

empcov = ic/nsims

list( empiricalcover = empcov )

}
```

VITA

ROMUAL E. TCHOUTA

Education:	B.S. Mathematics, University of Buea, Buea, Cameroon, 2005 M.S. Mathematics , East Tennessee State University Johnson City, Tennessee, 2008
Professional Experience:	Graduate Assistant, East Tennessee State University, Johnson City, Tennessee, 2006–2008