Methods for Estimating

Reference Intervals

METHODS FOR ESTIMATING REFERENCE INTERVALS

BY

CAITLIN H. DALY, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF McMASTER UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2014) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE: Methods for Estimating Reference Intervals

AUTHOR: Caitlin H. Daly

B.Sc., (Mathematics & Statistics)

McMaster University, Hamilton, Canada

SUPERVISOR: Dr. Jemila S. Hamid

NUMBER OF PAGES: ix, 102

Abstract

Reference intervals (RIs) are sets of percentiles that outline the range of laboratory test results belonging to healthy individuals. They are essential for the interpretation of laboratory test results. A wide variety of factors affect the validity of RIs. Among them are the statistical methods used to estimate RIs. However, little investigation has gone into the effect that different statistical methods have on the resulting RIs. This is particularly needed as the complexity of paediatric data makes it difficult to estimate RIs. These difficulties, however, can be addressed using appropriate statistical techniques, provided that there is an outline of scenarios under which these techniques are truly "appropriate".

The objective of this thesis is to provide a thorough investigation into the effect of different statistical methods on RIs. A systematic review was first conducted with a focus on paediatric RIs. The results of this review revealed that critical analysis steps are often overlooked due to complicated paediatric data. Even though a guideline addressing the establishment of RIs is available, there is great heterogeneity in the statistical methods chosen to estimate paediatric RIs.

An extensive simulation involving the three most commonly used approaches to estimate RIs (the parametric, non-parametric, and robust methods) was also conducted to

investigate and compare the performance of the different methods. The simulation results show that, when data follows a Gaussian distribution, or close to it, the parametric method provides the best estimates. The non-parametric method did not provide the best estimates of RIs (compared to the parametric method) unless data was highly skewed and/or large sample sizes were used.

In addition, the bias and MSE associated with the parametric method when data follows a Gaussian distribution was mathematically derived, which may lead to the development of a bias corrected and more precise approach in the future.

Acknowledgements

I would like to thank my wonderful supervisor, Dr. Jemila Hamid for her exceptional guidance and much appreciated support throughout my Master's studies. I am very grateful to have her as a role model and for the experience I have acquired while under her supervision. I would also like to thank my supervisory committee members Dr. Gregory Pond and Dr. Roman Viveros-Aguilera for their enthusiasm and invaluable input into my research. I am especially grateful for their ideas, which have provided me direction for my future work.

A special thank you to the CALIPER team at the Hospital for Sick Children. This thesis was funded by CIHR through the CALIPER team, and the topic was inspired by my experience as a co-op student for the CALIPER team in 2011. Drs. Adeli, Colantonio and Kyriakopoulou have all played a pivotal role in motivating me to pursue a career in research.

To Jarred and my friends inside and outside of school, thank you for the encouragement and much needed moments of fun. Last, but not least, I would like to thank my amazing parents and family for their endless love and support. I always look forward to the warm welcome home after a long day at school and am grateful to have my parents to talk to at any time of the day.

Contents

List of Figures

Chapter 1

Introduction

When an individual presents symptoms that affect their day-to-day activities, they may consult a medical professional to diagnose and treat the cause of these symptoms. The diagnosis of a potential illness not only requires careful examination of the patient at hand, but also good understanding of individuals presenting "normal" indicators of health. Knowing what is "normal" helps clinicians in recognizing abnormalities within a patient's health, investigating them, and reaching an informed diagnosis that will help facilitate effective treatment and disease management. However, accurately defining what is "normal" is a difficult task, and carries a great responsibility.

Reference intervals (RIs) provide clinicians a normal range for comparison when evaluating and interpreting a patient's laboratory test results. Current guidelines for establishing RIs provided by the Clinical Laboratory Standards Institute (CLSI) define RIs as ranges of values within which a specified percentage of measurements from healthy individuals would fall (Clinical and Laboratory Standards Institute [CLSI], 2008). When clinicians refer to RIs, they rely on the assumption that these RIs properly reflect the range of results expected from healthy individuals sharing similar characteristics to a

patient in question. If the patient's test results do not fall within the distribution of healthy individuals' results, then this will raise alarm in regards to the patient's health status. Thus, proper analysis of measurements from healthy individuals leading to the establishment of RIs is crucial for the proper interpretation of a patient's results.

Traditionally, a RI is established using the central 95% range of measurements defined by the 2.5^{th} and 97.5^{th} rank percentiles, as recommended by the CLSI guideline (CLSI, 2008). To interpret a patient's laboratory test result using a RI, a clinician simply has to verify whether the patient's test result falls within the RI or not. If the result falls within the RI, then the result is considered "normal". If the result falls outside the RI, then the result is considered "abnormal" and a clinician may relate this abnormality to a known illness, or carry out further tests to determine the cause of the abnormality. For example, suppose a 16-year-old male presents symptoms of a kidney disease to a physician. The physician may send the male to a laboratory with a requisition for blood analysis, with creatinine highlighted as an analyte to be tested. Creatinine is an indicator of renal function (Daniels, 2010). The physician receives a report for the male's blood work from the laboratory, which indicates the male's creatinine level as 112 μmol/L. This result is accompanied by a 95% RI for 15 to <19-year-old males: 55.1 – 95.5 μmol/L (Colantonio et al., 2012). Since 112 μmol/L is outside this RI, the physician recognizes this result as abnormal and proceeds with his/her diagnosis for the male.

A laboratory that analyzes patient samples should accompany test results with RIs that are specific to its own practices. To establish RIs anew, a laboratory must collect

samples from healthy individuals within its reference population, analyze the samples using its own instrumentation, and then use the appropriate statistical methods to compute the RIs (CLSI, 2008). For example, the Canadian Laboratory Initiative in Paediatric Reference Intervals (CALIPER) group at the Hospital for Sick Children in Toronto, Canada established a database of paediatric RIs using samples from healthy children from birth to 18 years of age living in the greater Toronto area that were analyzed by the Abbott ARCHITECT c8000 analyzer (Colantonio et al., 2012). These RIs may be used for comparison with samples from children living in the greater Toronto area that were analyzed by the Hospital for Sick Children using the Abbott ARCHITECT c8000 analyzer.

A wide variety of factors affect the validity of RIs. These include analytical factors such as instrumentation, reference population, sampling strategy, sample size, gender, age, and other demographic and lifestyle factors. The statistical methods used to construct RIs also play a major role on the RIs. However, they are often overlooked as factors that might have a considerable effect on the validity of the RIs. If suitable methods are not utilized given the type of data and sample size available, then a RI may deviate from the "true" RI for the population, leading to false conclusions of a patient's test results, which results in missed (from false negatives) or unnecessary (from false positives) treatment. The CLSI guideline was provided to address the need for a standardized statistical approach in selecting which methods to use. However, the methods suggested are not practicable for some laboratories due to the nature of their reference populations. In

particular, measurements from paediatric populations require extensive analysis due to the complex nature of paediatric data.

Children are continuously growing from birth to adolescence, and hence more age partitions than adult populations are often required. In addition to the standard covariates for adult populations (e.g., age, body mass index (BMI), gender, and/or ethnicity), maturity markers such as Tanner stage can greatly influence the composition of paediatric populations for which RIs should be provided. To capture all of these factors and ensure RIs are applicable, several partitions are warranted in paediatric data.

Another important consideration when establishing paediatric RIs is achieving sufficient sample size for every partition. This is particularly challenging in paediatric populations. Children are smaller than adults and thus blood procurement can be difficult. For example, 10 ml of blood could constitute to 10% of blood volume in a baby (Green et al., 2003). In addition to Research Ethics Board (REB) constraints, parental consents and costs make this a very demanding task.

Several national projects such as CALIPER, Children's Health Improvement Through Laboratory Diagnostics (CHILDx), German Health Interview and Examination Survey for Children and Adolescents (KiGGS), and Lifestyle Of Our Kids (LOOK) are currently underway to address the issue of outdated and unreliable paediatric RIs published in the past (CALIPER, 2014; Pediatric Reference Intervals, 2014; KiGGS, 2014; LOOK Lifestyle Study, 2014). With these projects underway, it is very important now more than ever to develop an outline of circumstantial methods in order to avoid the

unnecessary variability that may exist between these groups' published RIs, strictly due to differences in statistical methodology. Minimizing the differences between methodologies and providing a unified framework for selecting appropriate statistical methods will help clinicians compare RIs that are produced by various studies and identify any differences that may exist between populations. To make this possible, a thorough investigation is required to determine the impact various statistical methods have on resulting RIs.

The objective of this thesis is divided into two major components. First, a systematic review of medical literature was conducted with the aim of 1) investigating current literature on paediatric RIs with a focus on statistical methods that are used to construct paediatric-specific RIs, and 2) identifying gaps in the choice and implementation of the methods and reporting of the results. Details on the approach used and the results of the review are provided in Chapter 3. This work has also been published in Clinical Biochemistry (Daly et al., 2013).

The second objective of this thesis is to empirically evaluate and compare the performances of the different methods of RI estimation that are currently being used in practice. To this effect, extensive simulations were conducted using scenarios similar to real datasets across many sample sizes using Gaussian and skew normal distributions. The results of the simulation are provided in Chapter 4. Illustrations using real datasets are also provided in Chapter 5. In Chapter 2, the different methods used to estimate RIs are provided. Chapter 6 provides a discussion of the results presented in this thesis.

Chapter 2

Methodology

2.1 Background

Once reference data from a healthy population is made available, there are three key stages involved in the construction of reference intervals (RIs). In most cases, the first step is outlier detection. Outliers must be removed from data, even when using samples from a healthy population, as there is no guarantee that a "healthy" individual does not have an underlying disease that has not been diagnosed or detected.

There are several methods available to detect outliers. However, a select few pertain to RIs, as it is not unusual for analytes to follow a non-Gaussian distribution. Two of the most common methods used in the development of RIs are the Dixon method with a criterion proposed by Reed et al. and the Tukey method (Dixon, 1953; Reed et al., 1971; Tukey, 1977). Both of these methods have been recommended in the Clinical Laboratory Standard Institution (CLSI) guideline (CLSI, 2008). The Dixon method compares the distance between a suspected outlier and its neighbour to a proportion of the distance between the suspected outlier and the opposite endpoint of the data. It can only be applied

to one suspected outlier at one time. If it is in the right tail, the following criterion is used:

$$x_{(i)} - x_{(i-1)} > \frac{1}{3}\left(x_{(i)} - x_{(1)}\right)$$

where $x_{(i)}$ can either be the largest suspected outlier or the smallest of a suspected group of large outliers. Alternatively, if the suspected outlier is in the left tail, the following criterion is used:

$$x_{(i+1)} - x_{(i)} > \frac{1}{3}\left(x_{(n)} - x_{(i)}\right)$$

where, $x_{(i)}$ can either be the smallest suspected outlier or the largest of a suspected group of small outliers. If the appropriate criterion is satisfied for a suspected outlier, then the outlier, along with any suspected outliers above or below it should be removed, depending on whether the outliers lie in the right or left tail, respectively. The Tukey method, on the other hand, considers outliers in both tails simultaneously. Any observations smaller than the minimum (min) and or larger than the maximum (max) values given below should be removed:

$$\min = \mathrm{Q1} - 1.5\mathrm{IQR}; \max = \mathrm{Q3} + 1.5\mathrm{IQR}$$

where, Q1 is the first quartile of the observations, Q3 is the third quartile, and IQR is the interquartile range, which is defined as the difference between Q1 and Q3. Horn et al. recommends that data should be transformed to a Gaussian distribution before conducting this test (Horn et al., 2005).

After screening for outliers, data is partitioned into homogenous groups to reflect the changes that occur with various biological parameters. Partitions are generally determined

visually, often subjectively, as there is no known way to automatically compute them. Traditionally, age and gender are key covariates considered in partitioning data. However, with many countries becoming more culturally diverse, ethnicity should not go unnoticed. In addition, since many populations are experiencing increasing weight and obesity trends, weight factors such as BMI should also be considered. More specifically, for the paediatric population, developmental/Tanner stages that mark the progress of puberty are imperative for some analytes.

Although outlier detection is often performed at the initial stage, it is advisable to check for outliers after partitioning is done. This is primarily due to the fact that some observations may not appear to be extreme values among the entire dataset, but could appear to be much larger or smaller than most of the values in their partitioned group.

A common way to determine partitions is through categorical intervals. For example, for age, RIs can be calculated for every one-year interval, for height, ten-inch intervals. Another way to determine partitions is through previous clinical knowledge. If various stages throughout child development are known to affect a particular analyte, then biological covariates can be divided into well-known groups to reflect these stages. Finally, partitions can be determined through visual inspection of data. Plotting box-plots or parameters such as mean and standard deviation for each age and gender, or any other covariate of interest, can provide some insight as to which covariate values share similar compositions of analyte values.

After creating initial partitions through one of the three methods mentioned above, each partition should be tested against subsequent partitions to confirm whether these partitions should remain separate, or be combined. The CLSI guideline suggests several approaches proposed by Harris et al., Lahti et al., and Sinton et al. in performing these tests (Harris et al., 1990; Lahti et al., 2002; Sinton et al., 1986). The method proposed by Harris et al. evaluates the difference between the means and spreads of the two groups. That is, if at least one of the following two criterions are satisfied:

$$z_{calc} > z_{crit} \qquad \frac{s_2}{s_2 - s_1} < 3$$

where, $z_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, $z_{crit} = 3\sqrt{\frac{n_1 + n_2}{240}}$, $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the two groups in question, s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes, then the two groups should be kept separate. For this method, both groups under question must follow a Gaussian distribution. If not, this may be achieved by using a transformation.

Lahti et al.'s method also requires that the two groups in question follow a Gaussian distribution. This method involves the preliminary calculation of RIs for the two groups together as one. In the case where the central 95% range of the distribution is taken as the RI, the 2.5th and 97.5th percentiles should be calculated for the two groups combined. The percentage of values lying outside of these intervals for each group in question should be

determined. If one group's percentage exceeds 4.1% while the other's is less than 0.9%, then partitioning is recommended.

Similarly, Sinton et al.'s approach first calculates the 95% RI for the two groups combined. Then, the difference between the means of the two groups is determined. If the difference between the means is at least 25% as large as the difference between the two reference limits of the combined groups, partitioning is warranted. Skewness is not permitted for the calculation of this method.

Various hypothesis tests can be also used as alternatives to these three methods. These include t-tests or one-way analysis of variance (ANOVA) accompanied with pairwise t-tests and Bonferroni's correction to determine differences between the means of two or more groups (Dunn, 1961). The t-test requires that the two populations follow the Gaussian distribution and each group is sampled independently of each other from their respective populations. One-way ANOVA requires the assumption that the population variances are equal, in addition to the assumptions of the t-test. The Mann-Whitney U-Test (or Wilcoxon Rank Sum Test) can be used as an alternative to the t-test when the data does not follow a Gaussian distribution to test for differences in variances (Mann et al., 1947).

2.2 Reference Interval Estimation

Once data is checked for outliers and partitioning is appropriately done with respect to relevant factors, 95% RIs are estimated for each partition, and confidence intervals (typically 90%) for the estimated limits of the RI are provided. This involves estimating the 2.5th and 97.5th percentiles of the data, which are also commonly referred to as the lower and upper limits of the RI. There are three main approaches currently used to calculate RIs. They either employ parametric or non-parametric methods, or involve a bootstrapping technique; among them, the non-parametric approach is the most commonly used. The parametric method requires the data to follow a Gaussian distribution, whereas the other two approaches do not.

2.2.1 Parametric Method of Estimating RIs

A theoretical 100(1-α)% RI (θ) can be calculated for a population from a Gaussian distribution using

$$\begin{aligned}\theta &= \left(\theta_L, \theta_U\right) \\ &= \left(\mu - z_{\alpha/2}\sigma,\ \mu + z_{\alpha/2}\sigma\right),\end{aligned}$$

where, θ_L and θ_U are the lower and upper limits of the RI, respectively, μ is the known mean of the Gaussian distribution, $z_{\alpha/2}$ is the upper $\frac{\alpha^{\text{th}}}{2}$ percentile from a standard normal distribution, and σ is the known standard deviation of the Gaussian distribution.

A 100(1-α)% RI ($\hat{\theta}$) for a dataset can be estimated using

$$\hat{\theta} = \left(\hat{\theta}_L, \hat{\theta}_U\right)$$
$$= \left(\bar{x} - z_{\alpha/2}s, \bar{x} + z_{\alpha/2}s\right)$$

where, $\bar{x}$ is the sample mean and s is the sample standard deviation (Soldberg, 2006). This method requires the data to have an underlying Gaussian distribution. Non-Gaussian data may be transformed to an approximate Gaussian distribution to permit the application of this method.

Confidence intervals for RIs estimated by the parametric method are not readily available in RI literature and are not mentioned in the CLSI guideline. Confidence intervals for each limit, however, can be derived as follows. Let $x_1,\ldots,x_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. A 100(1-c)% confidence interval for the lower limit an be approximately calculated as

$$\hat{\theta}_L \pm z_{c/2}se\left(\hat{\theta}_L\right)$$

where,

$$\hat{\theta}_L \pm z_{c/2}se\left(\hat{\theta}_L\right) = \bar{x} - z_{\alpha/2}s \pm z_{c/2}se\left(\bar{x} - z_{\alpha/2}s\right) \quad \text{and}$$

$$se\left(\bar{x} - z_{\alpha/2}s\right) = \sqrt{Var\left(\bar{x} - z_{\alpha/2}s\right)}$$
$$= \sqrt{Var\left(\bar{x}\right) + z_{\alpha/2}^2 Var\left(s\right)}$$
$$= \sqrt{Var\left(\bar{x}\right) + z_{\alpha/2}^2 Var\left(\frac{\sigma}{\sqrt{n-1}}\frac{\sqrt{n-1}}{\sigma}s\right)}$$

$$= \sqrt{Var\left(\bar{x}\right) + z_{\alpha/2}^2 \frac{\sigma^2}{n-1} Var\left(\frac{\sqrt{n-1}}{\sigma} s\right)}.$$

Note that $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $\frac{\sqrt{n-1}}{\sigma} s \sim \chi_{n-1}$, where χ_{n-1} is the chi distribution. Hence,

$$se\left(\bar{x} - z_{\alpha/2} s\right) = \sqrt{\frac{\sigma^2}{n} + z_{\alpha/2}^2 \frac{\sigma^2}{n-1}\left[(n-1) - \left(\sqrt{2}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right]}$$

$$= \sqrt{\frac{\sigma^2}{n} + z_{\alpha/2}^2 \frac{\sigma^2}{n-1}\left[(n-1) - 2\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right]}.$$

We can estimate the standard error by using s^2 in place of σ^2. Thus,

$$\bar{x} - z_{\alpha/2} s \pm z_{c/2} \sqrt{\frac{s^2}{n} + z_{\alpha/2}^2 \frac{s^2}{n-1}\left[(n-1) - 2\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right]}$$

is a 100(1-c)% confidence interval for the lower limit of a 100(1-α)% RI estimated by the parametric method. Similarly,

$$\bar{x} + z_{\alpha/2} s \pm z_{c/2} \sqrt{\frac{s^2}{n} + z_{\alpha/2}^2 \frac{s^2}{n-1}\left[(n-1) - 2\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right]}$$

is a 100(1-c)% confidence interval for the upper limit of a 100(1-α)% RI estimated by the parametric method.

2.2.2 Non-Parametric Method of Estimating RIs

The non-parametric method is a simple way to calculate RIs empirically using ranks and does not require any assumptions regarding the distribution of the data (CLSI, 2008). To calculate the central 100(1-α)% RIs using the non-parametric method, the sample values for a given partition must be sorted from least to greatest, then ranked using whole numbers from 1 to n, where n is the size of the sample. Next, the following are calculated:

$$\mathrm{r}_1 = \frac{\alpha}{2}(n+1)$$

$$\mathrm{r}_2 = \left(1 - \frac{\alpha}{2}\right)(n+1).$$

The observations whose ranks correspond to r_1 and r_2 are then recorded as the $\frac{\alpha}{2}^{\text{th}}$ and $\left(1-\frac{\alpha}{2}\right)^{\text{th}}$ percentiles, respectively.

The CLSI guideline notes that a minimum sample size of $n = (100/\mathrm{P}) - 1$ is required to distinguish between percentiles that are P% apart (CLSI, 2008). Thus, when 95% RIs are desired, (requiring the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles to be distinguished from the 5$^{\text{th}}$ and 95$^{\text{th}}$, respectively), a minimum sample size of $n = (100/2.5) - 1 = 39$ is required.

Ranked observations are used to define confidence intervals for RIs established by the non-parametric rank method. The ranks of the observations that define a $100(1 -$

c)% confidence interval for the $\frac{\alpha}{2}^{\text{th}}$ percentile are determined by the highest integer a and smallest integer b such that

$$\sum_{i=a}^{b-1} \binom{n}{i} \left(\frac{\alpha}{2}\right)^{i} \left(1-\frac{\alpha}{2}\right)^{n-i} \geq 1-c.$$

The ranks of the observations that define a $100(1-c)\%$ confidence interval for the $\left(1-\frac{\propto}{2}\right)^{\text{th}}$ percentile are determined by $y = n - b + 1$ and $z = n - a + 1$.

2.2.3 Robust Method of Estimating RIs

The coined "robust method" uses an iterative process to compute a measure of the center of a dataset, denoted by T_{bi}, to derive 100(1-α)% RIs (Horn et al., 2005). First, T_{bi} is taken to be the median ($\tilde{x}$) of the n observations ($x_i, i = 1, \dots, n$) of a dataset and the median absolute deviation (MAD) is calculated using:

$$MAD = median(|x_i - \tilde{x}|).$$

Then, the observations of the dataset are weighted using

$$w_i = \begin{cases} x_i(1-x_i^2)^2 & \text{if } |u_i| < 1 \\ 0 & otherwise \end{cases}$$

where,

$$u_i = \frac{x_i - T_{bi}}{3.7\left(\frac{MAD}{0.6745}\right)}.$$

Note that 3.7 is a tuning constant chosen to permit the estimation of T_{bi} with minimal impact from outliers, and 0.6745 is chosen since $\sigma \approx \frac{MAD}{0.6745}$ when data is from a $N(\mu, \sigma^2)$ distribution. For further details, refer to (Horn et al., 2005). T_{bi} is then updated with the weighted data using

$$T_{bi} = \frac{\sum_{i=1}^{n} \boldsymbol{w_i x_i}}{\sum_{i=1}^{n} \boldsymbol{w_i}}.$$

The weighted observations are then reweighted using the updated value of T_{bi}, and the updated value of T_{bi} is once again updated using the above calculations. This process is repeated until the change in consecutive updated values of T_{bi} is insignificant. The 100(1-α)% RIs are then computed using

$$T_{bi} \pm t_{n-1,\propto/2}\sqrt{S_T^2[3.7] + s_{bi}^2[c_2]}$$

where, $t_{n-1,\propto/2}$ is the upper $\frac{\alpha^{\text{th}}}{2}$ percentile from a Student's t-distribution with $(n-1)$ degrees of freedom,

$$c_2 = \frac{1}{0.5813 - 0.607227(1-\propto)}, \text{ where, } 0.05 \leq \alpha \leq 0.50,$$

$$S_T[3.7] = 3.7(s_{bi}[3.7])\sqrt{\frac{\sum_{-1<v_i<1} v_i^2(1-v_i^2)^4}{\left(\sum_{-1<v_i<1}(1-v_i^2)(1-5v_i^2)\right)\max\left(1,-1+\sum_{-1<v_i<1}(1-v_i^2)(1-5v_i^2)\right)}},$$

where, $v_i = \frac{x_i - T_{bi}}{3.7(s_{bi}[3.7])}$,

$$s_{bi}[3.7] = 3.7s\sqrt{\frac{n\sum_{-1<w_i<1} w_i^2(1-w_i^2)^4}{\left(\sum_{-1<w_i<1}(1-w_i^2)(1-5w_i^2)\right)\max\left(1,-1+\sum_{-1<w_i<1}(1-w_i^2)(1-5w_i^2)\right)}},$$

where, $w_i = \frac{x_i - \tilde{x}}{3.7s}$ and $s = \frac{MAD}{0.6745}$, and

$$s_{bi}[c_2] = 205.6s \sqrt{\frac{n \sum_{-1<z_i<1} z_i^2 (1-z_i^2)^4}{\left(\sum_{-1<z_i<1} (1-z_i^2)(1-5z_i^2)\right) \max\left(1, -1 + \sum_{-1<z_i<1} (1-z_i^2)(1-5z_i^2)\right)}},$$

where, $z_i = \frac{x_i - \tilde{x}}{c_2 s}$ and $s = \frac{MAD}{0.6745}$. Note that 205.6 is a tuning constant chosen to capture the variability of the data. For further details, refer to (Horn et al., 2005).

Although a Gaussian distribution is not required for this method, Horn et al. recommends transformation to a more symmetric dataset to help better estimate RIs for skewed distributions (Horn et al., 2005). Percentile bootstrap estimates are used to construct $100(1-c)\%$ confidence intervals for RIs established using the robust method (Efron et al., 1993).

Chapter 3

Systematic Review of Methods in Paediatric RIs

3.1 Background

Systematic reviews are a commonly used approach in biomedial sciences. They are literature reviews that collect and screen articles presenting evidence that address a specific research question (Higgins et al., 2011). The systematic nature of these reviews minimizes the bias that may be introduced when researchers gather articles without a well defined plan. To conduct a systematic review, search terms relevant to the research question, along with inclusion and exclusion criteria must be predefined prior to conducting the literature search. In addition, at least two reviewers must screen relevant articles against the inclusion and exclusion criteria and agree upon which articles to include and exclude. This should return a wide range of relevant articles to review. The relevant articles are then assessed and key characterestics of each study are presented in a final report.

The first part of this thesis involves a systematic review of current literature with respect to statistical methods used to establish paediatric reference intervals (RIs). The review was conducted with the aim of 1) investigating current practice on paediatric RIs with a focus on statistical methods that are used to construct paediatric-specific RIs, 2) identifying gaps in the choice and implementation of the methods and reporting of the results, and 3) tailor the investigative work into the methodology behind the estimation of paediatric RIs so that weak areas of practice can be improved. This systematic review is published in Clinical Biochemistry and a portion of it is provided in the following sections of this chapter (Daly et al., 2013).

3.2 Search Criteria and Review Strategy

An electronic search on the Embase, MEDLINE and PubMed databases was conducted on May 28, 2012. Three themes were pre-identified as the search criteria. These themes were: “establishing”, “paediatric”, and “reference intervals”. Within each theme, a list of keywords or phrases was developed, which included various synonyms of the three themes commonly used in past literature. Effort was made in the search to ensure that some combination of the three themes was necessary for an article to be included in the search results. Search terms within each theme were combined with “OR”, and themes were combined with “AND”. Figure 3.1 represents the three themes, as well as the search words within each theme, that were used in the search of literature.

English articles published from January 1st, 2011 to May 18, 2012 were considered in this systematic review. Duplicated articles were deleted before the initial screening process. After removing duplicates, two reviewers were then presented with the resulting unique articles. The reviewers proceeded to independently review the title and abstracts of these articles against predetermined inclusion and exclusion criteria. Articles included in the systematic review presented new paediatric RIs established by the authors for intended public or in-house use. Articles were excluded if the RIs were calculated based on unhealthy samples, samples were outside the birth to less than 19 year age range, and/or samples were from a non-human population. In addition, if a study in the article used longitudinal data, strictly cited RIs for diagnostic or validation purposes, or simply did not establish new RIs, it was excluded. Following the exclusion of irrelevant articles, the same two reviewers screened the resulting articles' full text against the inclusion and exclusion criteria. If any disagreement arose during these processes, it was resolved either by discussion or consultation through a third reviewer. After reviewing all prospective articles, the reference lists of the included full-text articles were screened for additional relevant articles.

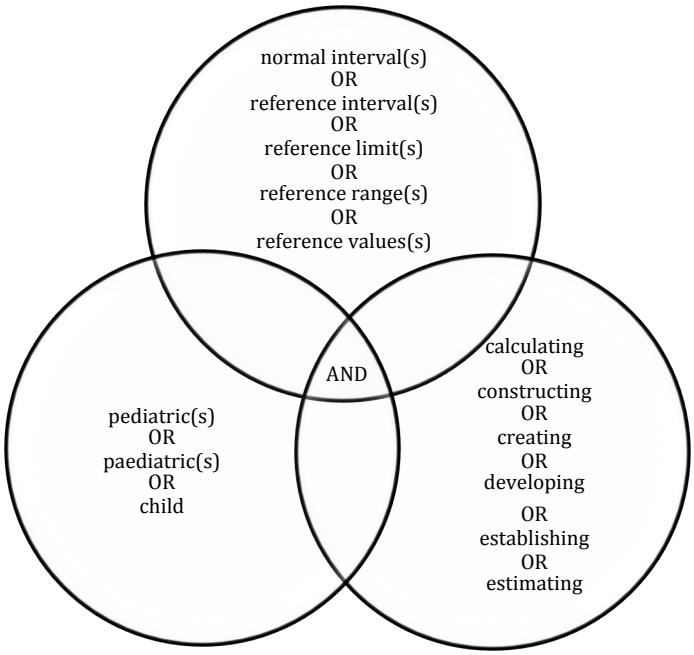

Figure 3.1: *Search criteria used in the systematic review.*

3.3 Results of Systematic Review

In total, 373 articles were initially returned from the keyword search, of which 195 were found to be unique. Through the initial screening of titles and abstracts, using the inclusion/exclusion criteria outlined in the Section 3.2, 37 articles were considered relevant. Following review of these articles' full texts, a final number of 22 articles were kept for the review. The reference lists of these articles were reviewed for any additional relevant articles. However, none were found. Figure 3.2 illustrates this process in more detail.

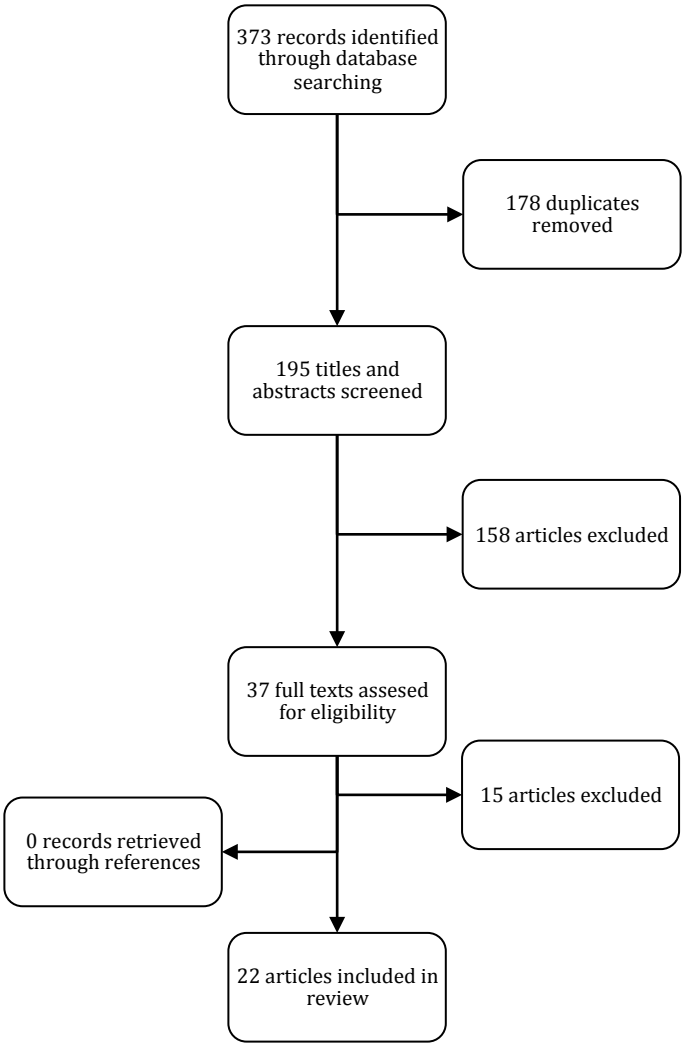

Figure 3.2: *Flow chart describing the review process.*

The distributions of the studies, with respect to the statistical methods used in the process of establishing RIs, are provided in Table 1. Among the articles included in our review, 59% stated that an outlier detection method was employed in the analysis of data. In terms of partitioning, the majority (about 96%) of the studies addressed variability with respect to covariates, identified the need for partitioning, and performed partitioning.

Even though the remaining articles (4%) investigated the influence various covariates had on the mean analyte values, they did not account for this in the estimation of RIs through partitioning.

Among the studies that performed partitioning, 24% of articles did not mention any application of statistical methods to test the appropriateness of partitions. Of the articles that performed partitioning, 9% created partitions based on clinical knowledge, 9% determined partitions visually, and 5% did not explicitly state the method they used to determine their partitions. The remainder used categorical partitions, that is, covariates were divided into equal, identifiable intervals. Of the studies that tested appropriateness of partitions, 31% did not collapse all insignificant partitions.

A large percentage of included articles (59%) employed some variation of methods recommended by the Clinical and Laboratory Standards Institute (CLSI) guideline to calculate RIs. However, only 32% of articles directly cited and followed the CLSI guideline, ensuring the use of appropriate methods when the minimum sample size of 120 was not met.

Table 1: *Distribution of studies included in the systematic review with respect to the statistical methods required in the process of establishing of RIs.*

	Percentage of Studies included in the review[a]
Used methods mentioned in CLSI guideline [b]	59.1%
Directly made a reference to and followed the recommendations provided in CLSI guideline	31.8%
Performed Partitioning	95.5%
Tested for differences among the partitions	76.2%
Collapsed insignificant partitions	68.8%
Checked for outliers	59.1%
Checked two way outlier detection[c]	15.4%
Statistical methods used	
Non-parametric	31.8%
Appropriately[d]	85.7%
Inappropriately[d]	14.3%
Parametric	22.7%
Robust	13.6%
Other[e]	40.9%
Reported confidence intervals	18.2%
For all	75%
For some	25%

[a] All percentages are out of the total number of studies included in the systematic review
[b] One or more of the methods used in the process to establish RIs is recognized by the CLSI guideline.
[c] Of those studies that detected outliers, consideration was given before and after partitioning.
[d] The non-parametric method was used appropriately when there was a minimum of 120 samples.
[e] Obtained limits from reference curves, or did not explicitly states how percentiles were derived.

Among the studies that utilized methods that are not strictly outlined in the CLSI guideline is the study by Uemera et al., where RIs for serum creatinine levels were established (Uemura et al., 2011). In this study, reference to the CLSI guideline was not made and it is not clear how outliers were detected or handled. However, the authors addressed the issue of gender, age and body size variability through partitioning. Age and body length were considered separately as partitioning criteria. That is, RIs were calculated with age as the partitioning factor and again with body length as the partitioning factor. In both cases, predetermined categorical intervals were used leading to large number of partitions. This might have led to small sample sizes for some of the partitions. The RIs were established using the 2.5^{th} and 97.5^{th} percentiles. However, there was no mention of how these percentiles were calculated (e.g., parametric vs. non-parametric). Moreover, similar to the majority of the articles included in this review, confidence intervals or any measurements of variability (e.g., standard error) were not given.

Many other articles used predefined categorical partitions, however, a select few articles included in this review conducted tests for differences within predefined categorical partitions (Adibi et al., 2012; Clifford et al., 2011; Goh et al., 2011; Tamimi et al., 2011; and Tamimi et al., 2012). In addition, Cinaz et al. and Hulecki et al. included regression analysis to explore possible significant covariates (Cinaz et al., 2012; Hulecki et al., 2011). However, statistically insignificant differences and/or significant covariates were not considered in the establishment of the final RIs as the categorical partitions had

already been determined. Nonetheless, the findings were stated in the papers for future consideration.

The majority of articles employed either the parametric (23%), non-parametric (32%), or robust (14%) methods to construct RIs. Of the articles that used the non-parametric approach, 14% of them did not meet the minimum sample size of 120 to estimate their RIs. Furthermore, an alarmingly small percentage of articles (18%) accompanied their RIs with confidence intervals.

3.4 Identified Gaps and Recommendations

There are a number of steps involved in the establishment of RIs, and in a paediatric setting, this process can become quite cumbersome as children are continuously growing from birth to adolescence. This systematic review has shown that the complexity of paediatric data is being acknowledged by many and some measures are being taken to address these problems. However, the review has also identified gaps in the establishment (estimation) and reporting of paediatric RIs, and there is still room for improvement. Moreover, guidelines specific to paediatric populations are needed to encourage appropriate and more standardized use of statistical methods.

The first step in the establishment of RIs is outlier detection and handling of the identified outliers. Removing outliers from data is a common practice, even when using samples from a healthy population, as there is no guarantee that a "healthy" individual does not have an underlying disease that has not been diagnosed/detected. However,

outliers may also tell us important information about a given analyte and may represent a natural variability within a given dataset. Hence, careful consideration must be taken before their removal. It is also helpful to perform sensitivity analysis to investigate the robustness of the RIs with respect to outlier observations.

There are several existing statistical methods available to detect outliers (Sen et al., 1990). However, a select few pertain to RIs. Two of the most commonly used outlier detection methods in the development of RIs are the Dixon method with a criterion proposed by Reed et al. and the Tukey method (Dixon, 1953; Reed et al., 1971;Tukey, 1977). Both of these methods have been recommended in the CLSI guideline (CLSI, 2008). However, it is not clear why these specific approaches were selected. We are not aware of any study done to formally assess their performance under different situations. For instance, both the Dixon and Tukey methods were developed under the assumption of a Gaussian distribution and may not be appropriate for non-Gaussian data. Even when data follows a Gaussian distribution, these approaches may not be optimal for small sample size. Studies are, therefore, needed to understand the behavior of these methods under many scenarios.

After screening for outliers, data is often partitioned into homogenous groups to reflect the changes that occur with various biological parameters. This is particularly important in paediatric populations, where more partitions (than adult populations) are required to reflect child development.

A common way to determine partitions is through categorical intervals. For example, for age, RIs can be calculated for every 1-year, 2-year or 5-year interval. Some studies included in this review produced RIs for categorical 1-year age intervals. However, this is not recommended. Not only does this create partitions that are not practical for clinical use, but also creates an unnecessarily large number of partitions leading to insufficient sample size for all or some of the partitions.

An alternative way of determining partitions is by visually investigating the data for possible homogenous groups. Visual examination of the data allows the researcher to identify groups that are similar in terms of their central location and spread of analyte values. This approach can be useful even when partitions have been established using pre-defined categories or clinical knowledge, allowing researchers to update their partitions based on the findings. Visual assessment can be performed by using simple scatter or box plots against age, gender, Tanner stage, ethnicity, or other relevant covariates to identify possible changes in the value of the analyte. This can be further improved by including a measure of variation, e.g., plot coefficient of variation instead of the mean.

Once partitions are made, different statistical methods can be used to compare the partitions and confirm whether they should remain separate or collapsed. This can be done using a simple sequential t-test or the Harris et al. method of comparing means and standard deviations for Gaussian data, and the Mann-Whitney U-Test (or Wilcoxon Rank Sum Test) for non-Gaussian (skewed) data (Harris et al., 1990; Mann et al., 1947). Alternatively, RIs can be calculated for all potential partitions and the resulting limits of

different groups are compared using the methods described by Lahti et al. and Sinton et al., as mentioned in the CLSI guideline (Lahti et al., 2002; Sinton et al., 1986). However, the guideline does not provide when, why and how to choose the optimal approach.

In addition, we would like to emphasize that testing partitions often involves performing multiple tests, leading to increasing false positive rates (type I error), leading to the conclusion that partitions are significantly different, when in fact they are not. To overcome this challenge, we recommend adjusting for multiple comparisons using methods such as the Bonferroni correction or the less conservative adjustment of Sidak (Dunn, 1961; Sidak, 1967). Note that, although outlier detection is traditionally conducted first when dealing with any type of dataset, we recommend the use of outlier detection and partitioning interchangeably. This combination of methods avoids the potential masking of outliers and false detection of significant differences or lack thereof between partitions or groups.

The large variation in methods used to establish RIs is a striking observation gathered from this systematic review. Some of the methods used by the studies considered in this review include: parametric, non-parametric, robust, parametric fractional polynomial curves, lambda-mu-sigma curves (referred to as the LMS method (Cole et al., 1992)), as well as some ad hoc approaches using the mean and standard deviation of the data.

When selecting a method to compute RIs, it is important to note that each method uses an underlying assumption and comes with its own advantages and disadvantages. In terms of the three main methods discussed in the CLSI guideline, the parametric approach

offers the advantage of more precise estimates over the non-parametric approach when the data follows a Gaussian distribution. This approach also requires less sample size than the non-parametric approach. However, the CLSI guideline does not state how much sample size is enough for the parametric approach and how the distributional assumptions (Gaussian) should be investigated. We strongly encourage researchers to use normal probability plots to investigate the distribution of the data and assess skewness before deciding to use (or not to use) the parametric approach.

When the data does not follow a Gaussian distribution, the CLSI guideline suggests the non-parametric method as a preferable approach. The guideline recommends a minimum sample size of 120 in order to calculate confidence intervals using the non-parametric approach (CLSI, 2008). When this condition is not met, the robust method is recommended.

In addition to methodological gaps, this review has also revealed gaps in the reporting of RIs. The majority of studies included in this review did not include confidence intervals for the reference limits and hence the RIs may not be reliable. Comparing the sizes of the confidence intervals to the sizes of the RIs provides some insight as to how precise (good) a RI is. For instance, it has been recommended that the width of a 90% confidence interval should be less than 0.2 times the width of a 95% RI (Harris et al., 1990). Consequently, if the width of a confidence interval is clinically unacceptable, then the estimate is not reliable. Moreover, the width of confidence intervals for each of the limits is a measure of the precision of the estimates and provides

information on the sampling variability of the RIs. Furthermore, confidence intervals (or standard error of) the reference limits can be used to judge the adequacy of the sample size being used in the analysis. For instance, a larger sample size may be required for a particular partition in order to increase the precision of the reference limits and hence improve the reliability of the RI.

The lack of confidence intervals in literature could perhaps be due to the absence of a method for calculating confidence intervals for the parametric method, and/or the difficulty of collecting sufficient sample size (as recommended by the CLSI guideline) for the non-parametric method. We recommend use of the bootstrap methods, where confidence intervals are calculated by re-sampling from the data (Efron, 1982). We would like to point out that, due to recent advances in computational power, bootstrapping and other re-sampling approaches are becoming more feasible for use by many medical researchers to calculate confidence intervals for estimates. These approaches are currently available in most statistical software packages and researchers providing RIs are encouraged to calculate confidence intervals using these approaches.

Several useful recommendations are made in the CLSI guideline, including which statistical methods to use to detect outliers and develop RIs, as well as recommendations regarding how much sample size is required. However, limited work has been done to study the impact the statistical methods have on the resulting RIs. Recently, a group from Utah performed an empirical comparison of three methods (parametric, transformed parametric, and bootstrap) using data from an adult population (Pavlov et al., 2012). Their

study highlights the advantages and shortcomings of the three methods in given scenarios. The scenarios include small or large sample sizes as well as data from analytes with Gaussian or non-Gaussian distributions. However, the study did not consider the most commonly used non-parametric approach, nor did it include the robust method to investigate its performance for small samples, and only used one outlier detection method. In addition, the empirical comparison was limited to adult populations and did not consider partitioning.

Moving forward, it would be beneficial to perform similar studies with more extensive coverage of statistical methods used in the RI estimation process. Such studies should also include an extensive simulation to assess performance of each method under different scenarios. Simulations allow researchers to compare estimates with the true values, which means a quantitative measure can be obtained to reflect the precision of a method. This would enable the development of a paediatric-specific guideline, which would help research groups calculate paediatric RIs using an appropriate and more standardized fashion. To this effect, a simulation study investigating the performance of three most commonly used approaches, the parametric, non-parametric, and robust methods, is presented in Chapter 4.

Chapter 4

Simulation

4.1 Motivation of Simulation

The systematic review presented in the previous chapter revealed two main statistical issues that can be addressed immediately (Daly et al., 2013). One is the selection of the method for reference interval (RI) estimation. Several methods were used in addition to the non-parametric and robust methods highlighted in the Clinical Laboratory Standards Institute (CLSI) guideline. However, the parametric, non-parametric, and robust methods are the most popular approaches being used. There is still some hesitancy regarding the selection of these methods based on the characteristics of the data available (such as normality and sample size). Therefore, there is a need to develop a guideline that covers the various scenarios related to the assumption of normality and sample size that arise in paediatric data and explicitly deliver recommendations pertaining to these scenarios.

The second issue that can be addressed immediately is the lack of confidence intervals or measure of precision of the RI estimates. This may be because of absence of literature describing confidence intervals for the parametric method, no alternative to

producing confidence intervals for non-parametric RIs with sample sizes less than 120, or the complex appearance of the robust method that drives laboratories away from using it. A confidence interval for RIs estimated by the parametric method is derived in this thesis and is presented in Chapter 2. We performed an extensive simulation to investigate the performance of the different methods of estimating RIs and the results are presented in this chapter.

4.2 Description of Simulation

Extensive simulations were performed to investigate the performances of the parametric, non-parametric, and robust methods for estimating RIs, where a total of 216 scenarios was generated, using combination of parameters provided in Table 2 below.

Table 2: *Mean, variance, sample size, and skewness used to generate data for simulation.*

	Distribution	
	Gaussian	Skew Normal
μ	2.45, 20, 74.82	2.45, 20, 74.82
σ^2	0.01, 9, 132.70	0.01, 9, 132.70
n	40, 80, 120, 160, 200, 240, 280, 320, 360, 400, 440, 480	40, 80, 120, 160, 200, 240, 280, 320, 360, 400, 440, 480
Skew	0	0.10, 0.25, 0.50, 0.75, 0.95

Essentially, data were generated from different distributions for each sample size. The mean, variance and skewness values were used to determine the distribution we were

investigating. The choice of parameters was motivated by real data, along with observed changes in bias and mean squared error (MSE) due to variance, not mean. As a result, small, moderate, and large values of variance were selected. The distribution with small variance ($\mu = 2.45, \sigma^2 = 0.01$) was generated to mimic the distribution of calcium values for females and males aged 1 to less than 19 years, collected by the Canadian Laboratory Initiative for Paediatric Reference Intervals (CALIPER) group (CALIPER, 2014). The distribution with large variance ($\mu = 74.82, \sigma^2 = 132.70$) was generated to mimic the distribution of creatinine (enzymatic) values for males aged 15 to less than 19 years, collected by the CALIPER group (CALIPER, 2014). The distribution with moderate variance ($\mu = 20, \sigma^2 = 9$) was generated to permit the examination of the performances of the methods with respect to variance increase. The variance of this distribution was specifically chosen as a value between the variances of the first and second distributions just mentioned.

For each scenario, 1000 datasets were generated from which RIs and their corresponding confidence intervals were computed using the parametric, non-parametric, and robust methods. The bias, $\text{Bias}\left(\hat{\theta}_j, \theta_j\right) = \text{E}\left(\hat{\theta}_j - \theta_j\right)$, and MSE, $\text{MSE}\left(\hat{\theta}_j, \theta_j\right) = \text{Var}\left(\hat{\theta}_j\right) + \left(\text{Bias}\left(\hat{\theta}_j, \theta_j\right)\right)^2$, for each limit were then empirically estimated, where, $\hat{\theta}_j$ is the estimated reference limit ($j = L, U$ for the lower and upper limits, respectively) of the RI and θ_j is the corresponding limit of the true RI. The limits of the

true RI were computed as the theoretical percentiles of the distribution under consideration. This was repeated 50 times to generate 50 estimates of bias and MSE, and the results were averaged over the 50 repetitions.

4.3 Calculation of Parameters for the Skew Normal Distributions

For simulation scenarios involving skewed distributions, the parameters of the skew normal distributions, for comparison purposes, were calculated so that the scenarios had the same mean and variance as the Gaussian scenarios. This was done as follows.

Consider the density of the skewed normal distribution

$$\varphi(z;\alpha) = 2\phi(z)\Phi(\alpha z), \quad -\infty < z < \infty,$$

denoted by *SN(*α), where ϕ and Φ are the density and distribution functions of a standard normal random variable, respectively (Azzalini, 2005). If Z~*SN*(α) and $Y = \xi + \omega Z$, where, $\xi \in R$, $\omega \in \mathbf{R}^+$, then Y~*SN*(ξ,ω^2,α), where ξ is the location parameter and ω is the shape parameter (Azzalini, 2005).

Suppose X~ *N*(μ, σ^2) and Y~*SN*(ξ,ω^2,α). Note the following mean, variance, skewness, and kurtosis of the random variable Y:

$$\mathrm{E}(Y) = \xi + \omega\sqrt{\frac{2}{\pi}}\delta, \quad \mathrm{Var}(Y) = \omega^2\left(1 - 2\frac{\delta^2}{\pi}\right),$$

$$\gamma_1 = \frac{4-\pi}{2}\frac{\left(\sqrt{\frac{2}{\pi}}\delta\right)^3}{\left(1-2\frac{\delta^2}{\pi}\right)^{3/2}}, \text{ and } \gamma_2 = 2(\pi-3)\frac{\left(\sqrt{\frac{2}{\pi}}\delta\right)^4}{\left(1-2\frac{\delta^2}{\pi}\right)^2}$$

where, $\delta = \frac{\alpha}{\sqrt{1+\alpha^2}}$.

These characteristics can be used to solve for ξ, ω, α, and δ such that

$$\mu = \xi + \omega\sqrt{\frac{2}{\pi}}\delta, \qquad \sigma^2 = \omega^2\left(1-2\frac{\delta^2}{\pi}\right) \qquad \text{and} \qquad \kappa = \frac{4-\pi}{2}\frac{\left(\sqrt{\frac{2}{\pi}}\delta\right)^3}{\left(1-2\frac{\delta^2}{\pi}\right)^{3/2}}$$

for a desired level of skewness (κ). The values of ξ, ω, α for a desired mean (μ), variance (σ^2) and skewness (κ) can be calculated with

$$|\delta| = \sqrt{\frac{\pi}{2}\frac{|\kappa|^{2/3}}{|\kappa|^{2/3}+\left(\frac{4-\pi}{2}\right)^{2/3}}}, \qquad \alpha = \frac{\delta}{\sqrt{1-\delta^2}}, \qquad \omega = \sqrt{\frac{\sigma^2}{1-\frac{2\delta^2}{\pi}}}, \qquad \text{and} \qquad \xi = \mu - \omega\delta\sqrt{\frac{2}{\pi}}.$$

4.4 Simulation Results

4.4.1 Results for the Gaussian Distribution

Three Gaussian distributions (N(2.45,0.01), N(20,9), N(74.82,132.70)) were considered, where 1000 datasets were generated from each of these distributions. This was executed using sample sizes of n = 40, 80, 120, 160, 200, 240, 280, 320, 360, 400, 440, and 480.

RIs were estimated for each dataset using the parametric, non-parametric, and robust methods (as described in Chapter 2). The bias and MSE of both the lower and upper limits were empirically estimated and recorded.

Results of the simulation for Gaussian data show that the parametric method produces lower limit estimates with mostly positive, but negligible, bias and upper limit estimates with mostly negative, but negligible, bias when sample size is small. However, as the sample size increases, negligible positive and negative biases are observed for both the lower and upper limit estimates. Plots of the bias of the lower and upper limit estimates for the parametric method are provided in Figure 4.1.

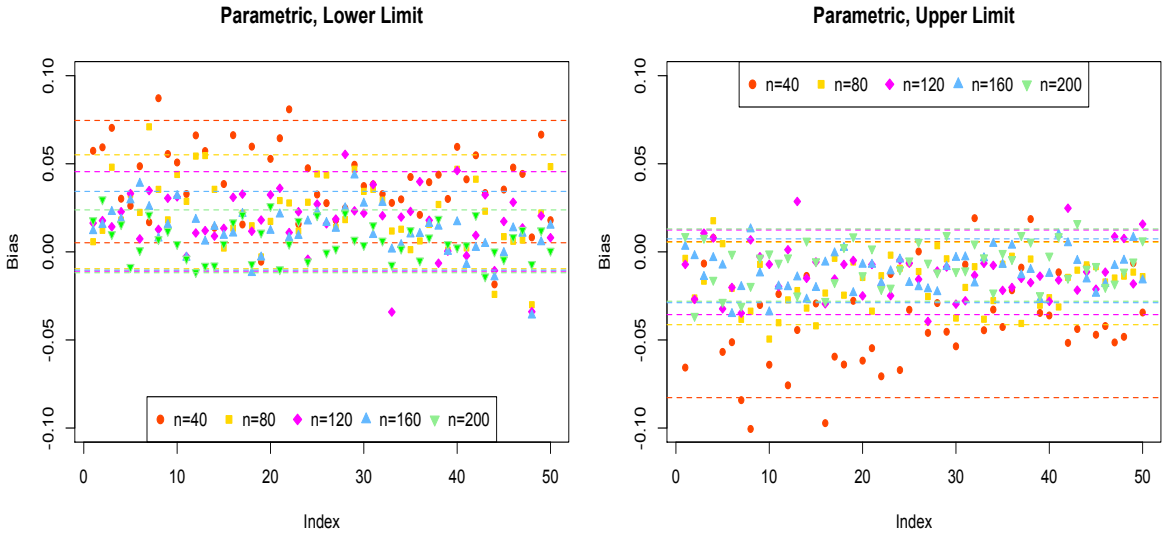

Figure 4.1: *Empirical bias, with corresponding 95% confidence intervals (indicated by dashed lines), for the parametric method, where data was generated from N(20,9).*

The magnitudes of the bias of the estimated lower limit and the estimated upper limit (Figure 4.1) are very close to 0. However, statistically there is a chance that laboratory

test results will be incorrectly identified as abnormal because they fall within the area below and near the estimated lower limit and within the area above and near the estimated upper limit. This chance is most likely non-existent clinically, since the value of the bias is negligible compared to the variability in the data. As an illustration, in Figure 4.2, the bias is so negligible that one cannot observe the difference between the true RI and the estimated RI.

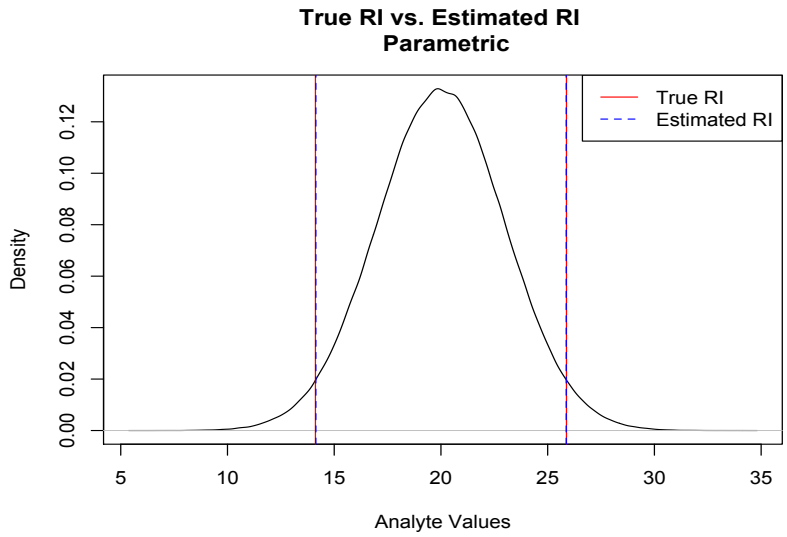

Figure 4.2: *True RI vs. estimated RI for the parametric method.*

In Figure 4.1, the spread of the bias decreases as sample size increases. This indicates that the variability of the bias of estimates for both the lower and upper limits is inversely proportional to sample size. In addition, as sample size increases, the distribution of bias for the estimates of both the lower and upper limits appears to center around 0, indicating that the estimates are asymptotically unbiased. In fact, we show on the following page that this indeed is the case mathematically.

Let $x_1,\ldots,x_n$ be a random sample from a $N\left(\mu,\sigma^2\right)$ distribution, and let $\bar{x}=\frac{1}{n}\sum_{n}^{1}x_i$ and $s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$. Then the bias of the lower limit $\left(\hat{\theta}_L\right)$ can be derived as follows:

$$\begin{aligned}\text{Bias}\left(\hat{\theta}_L,\theta_L\right)&=\text{E}\left(\hat{\theta}_L-\theta_L\right)\\&=\text{E}\left[\left(\bar{x}-z_{\alpha/2}s\right)-\left(\mu-z_{\alpha/2}\sigma\right)\right]\\&=\text{E}\left(\bar{x}-z_{\alpha/2}s\right)-\text{E}\left(\mu-z_{\alpha/2}\sigma\right)\\&=\text{E}\left(\bar{x}\right)-z_{\alpha/2}E\left(s\right)-\text{E}\left(\mu\right)+z_{\alpha/2}\text{E}\left(\sigma\right).\end{aligned}$$

Recall that $\bar{x}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$, consequently

$$\begin{aligned}\text{Bias}\left(\hat{\theta}_L,\theta_L\right)&=\mu-z_{\alpha/2}\text{E}\left(\frac{\sigma}{\sqrt{n-1}}\frac{\sqrt{n-1}}{\sigma}s\right)-\mu+z_{\alpha/2}\sigma\\&=-z_{\alpha/2}\frac{\sigma}{\sqrt{n-1}}\text{E}\left(\frac{\sqrt{n-1}}{\sigma}s\right)+z_{\alpha/2}\sigma.\end{aligned}$$

Moreover, $\frac{\sqrt{n-1}}{\sigma}s\sim\chi_{n-1}$. As a result,

$$\begin{aligned}\text{Bias}\left(\hat{\theta}_L,\theta_L\right)&=z_{\alpha/2}\sigma\left(1-\frac{1}{\sqrt{n-1}}\sqrt{2}\frac{\Gamma\left(\frac{\left(n-1\right)+1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)\\&=z_{\alpha/2}\sigma\left(1-\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right).\end{aligned}\qquad(1)$$

Similarly, the bias of the upper limit $\left(\hat{\theta}_U\right)$ can be derived as

$$\text{Bias}\left(\hat{\theta}_U, \theta_U\right) = z_{\alpha/2}\sigma\left(\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} - 1\right). \quad (2)$$

Now consider, $\dfrac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$, it can be re-written as (Graham et al., 1994)

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \approx \sqrt{\frac{n-1}{2}}\left(1 - \frac{1}{4(n-1)} + O\left(\frac{1}{n^2}\right)\right).$$

Therefore, $\lim_{n\to\infty}\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} = 1.$ (3)

This indicates that the bias of both the lower and upper limit estimates converge to 0. Thus, the estimates of the lower and upper limits produced by the parametric method are indeed asymptotically unbiased.

The non-parametric method, on the other hand, does not appear to produce unbiased estimates for the lower and upper limits. In fact, empirical results show that, as the sample size increases, bias for the non-parametric method converges to a constant different from zero. Plots of the bias of the lower and upper limit estimates for the non-parametric method are provided in Figure 4.3.

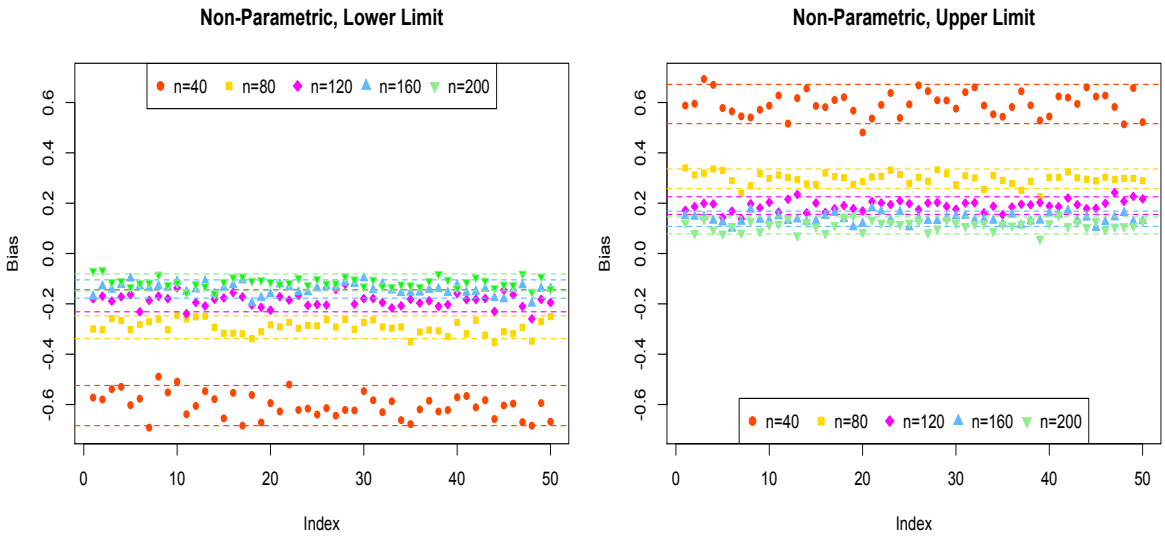

Figure 4.3: *Empirical bias, with corresponding 95% confidence intervals (indicated by dashed lines), for the non-parametric method, where data is generated from N(20,9).*

In general, the non-parametric method produces lower limit estimates with small, negative bias, and upper limit estimates with small, positive bias (Figure 4.3). Consequently, the estimated lower limit is smaller than the true lower limit for the population, and the estimated upper limit is larger than the true upper limit for the population (see Figure 4.4 for illustration). As a result, RIs estimated by the non-parametric method will be wider than the true RI for the population on average, leading to false negatives (missed treatment).

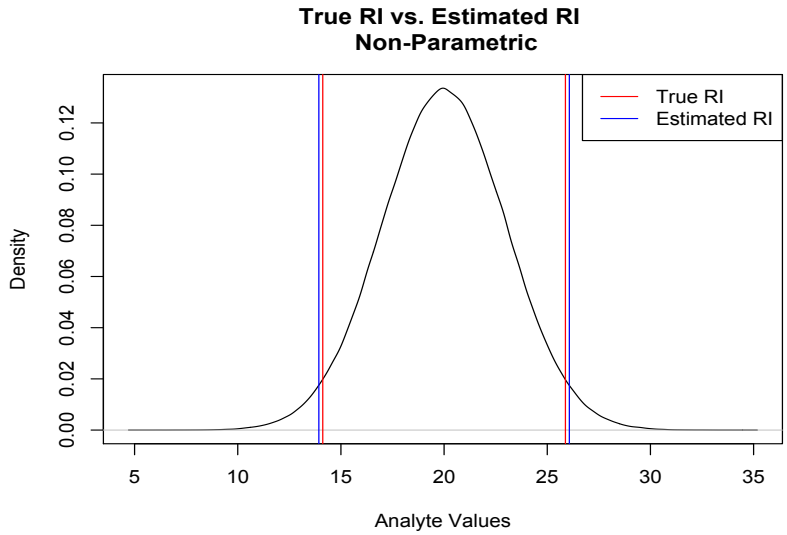

Figure 4.4: *True RI vs. estimated RI for non-parametric method.*

Compared to the parametric method, the bias for the non-parametric method is generally larger across all sample sizes (Table 3). However, like the parametric method, the variability of the bias of estimates for both the lower and upper limits obtained using the non-parametric method is inversely proportional to sample size (Figure 4.3). As sample size increases, the distribution of bias shifts towards 0 for both estimated limits. However, it converges to a value close to 0, indicating that the estimates are not asymptotically unbiased, unlike the parametric method. It is interesting to note the distinction between the distribution of bias when $n = 40$ and the distributions of bias for larger sample sizes (Figure 4.3). Although RIs estimated by the non-parametric method can be obtained with a sample size as little as 39, one should note the significant decrease in bias with an increase in sample size.

Table 3: *Average empirical bias for the three different methods, where data is generated from N(20,9).*

n	Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	0.0399	-0.0385	-0.6042	0.5940	-0.2677	0.2715
80	0.0228	-0.0179	-0.2927	0.2971	-0.1287	0.1351
120	0.0173	-0.0116	-0.1879	0.1903	-0.0828	0.0904
160	0.0120	-0.0107	-0.1410	0.1380	-0.0628	0.0660
200	0.0060	-0.0075	-0.1148	0.1115	-0.0540	0.0535
240	0.0071	-0.0037	-0.0939	0.0948	-0.0431	0.0470
280	0.0047	-0.0033	-0.0836	0.0827	-0.0381	0.0401
320	0.0046	-0.0041	-0.0677	0.0729	-0.0330	0.0339
360	0.0043	-0.0032	-0.0647	0.0617	-0.0292	0.0304
400	0.0039	-0.0036	-0.0566	0.0571	-0.0260	0.0268
440	0.0035	-0.0030	-0.0481	0.0531	-0.0243	0.0241
480	0.0025	-0.0033	-0.0477	0.0477	-0.0226	0.0219

The patterns of bias observed with the non-parametric method are similar to those observed with the robust method. Plots of the bias of the lower and upper limit estimates for the robust method are provided in Figure 4.5.

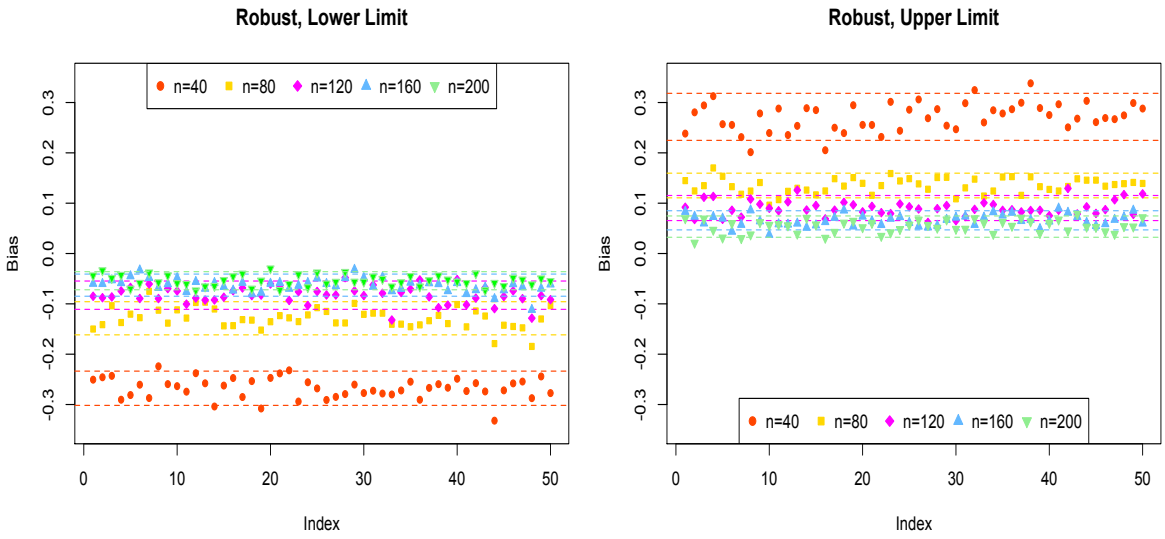

Figure 4.5: *Empirical bias, with corresponding 95% confidence intervals (indicated by dashed lines), for the robust method, where data is generated from N(20,9).*

The robust method produces lower limit estimates with small, negative bias, and upper limit estimates with small, positive bias (Figure 4.5). Consequently, the estimated lower limit is smaller than the true lower limit for the population, and the estimated upper limit is larger than the true upper limit for the population (see Figure 4.6 for illustration). As a result, RIs estimated by the robust method will be wider than the true RI for the population on average, leading to false negatives. Note that, although patterns in bias are quite similar between the non-parametric and robust methods, the robust method produces estimates with smaller bias than the non-parametric method (Table 3).

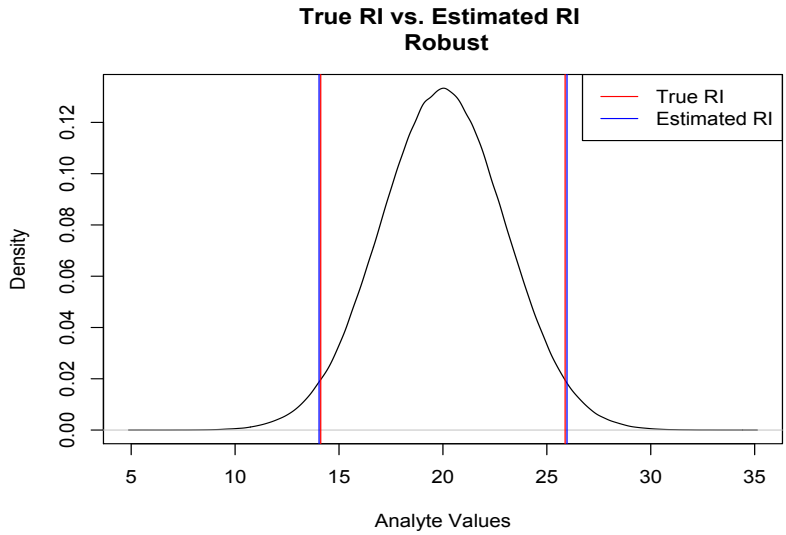

Figure 4.6: *True RI vs. estimated RI for robust method.*

Similar to the parametric and non-parametric methods, the variability of the bias of the lower and upper limit estimates produced by the robust method is inversely proportional to sample size (Figure 4.5). As sample size increases, the distribution of bias shifts towards 0 for both estimated limits. However, like the non-parametric approach, the estimates are not asymptotically unbiased. Estimates for the lower and upper limits appear to converge to negative and positive constants, respectively, as sample size increases. In addition, as with the non-parametric method, the distribution of bias for the robust method is distinct between $n = 40$ and larger sample sizes (Figure 4.5). Again, one should note the significant decrease in bias that accompanies an increase in sample size.

It is clear that the parametric method produces the least biased estimates, while the non-parametric method produces the most biased estimates, regardless of sample size. This is consistent regardless of the mean and variance of the distributions considered.

Note that the magnitude of the average bias for the estimated lower and upper limits is almost the same to two decimal places for all three methods (Table 3), which one would expect with symmetric distributions. As a result, the number of false positives associated with test results near the estimated lower limit will be about the same as the number of false positives associated with test results near the estimated upper limit, for limits estimated by the parametric method. The same is true for limits estimated by the non-parametric and robust methods, except with false negatives.

When mean squared error (MSE) is considered, we observed a significant difference between the performance of the non-parametric method and the performances of the parametric and robust methods. The parametric method provided uniformly lower MSE and the robust method slightly larger (almost negligible for some sample sizes) (Figure 4.7). However, the non-parametric method resulted in much larger MSE, especially when the variability of the data is large.

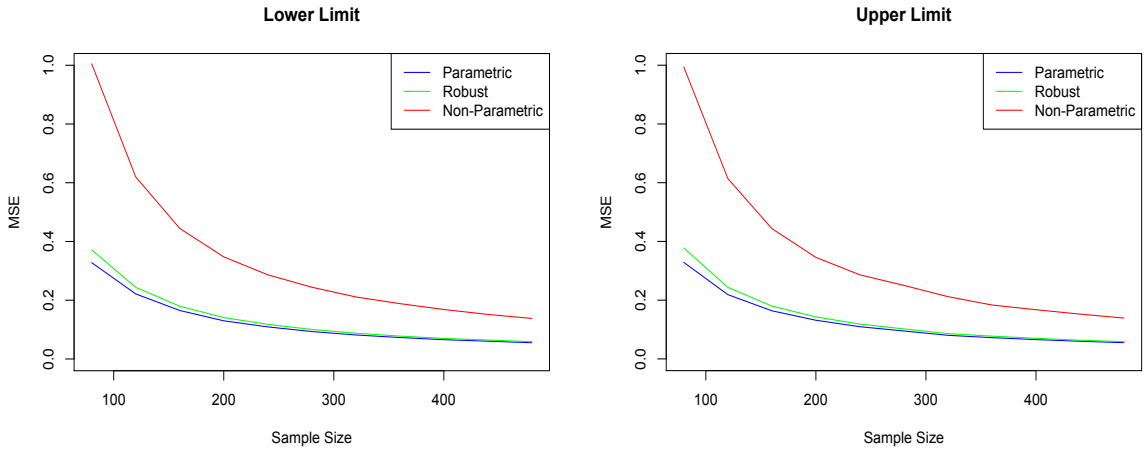

Figure 4.7: *Average empirical MSE for the three methods, where data is generated from N(20,9).*

For instance, for data from $N(20,9)$ and $n = 40$, the average MSE of the non-parametric lower and upper limit estimates is more than triple of that of the parametric estimates, and almost triple of that of the robust estimates (Table 4). Similarly, when $n = 120$, the average MSE of the non-parametric estimates (approximately 0.62 and 0.61) is almost triple of that of the parametric (approximately 0.22 and 0.22) and robust estimates (approximately 0.24 and 0.24). When $n = 480$, the largest sample size considered in this simulation study, the difference between the parametric and robust methods is negligible, but the MSE for the non-parametric method is still considerably larger.

Table 4: *Average empirical MSE for the three different methods, where data is generated from N(20,9).*

n	Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	0.6661	0.6655	2.4101	2.3821	0.8218	0.8259
80	0.3279	0.3286	1.0050	0.9941	0.3717	0.3777
120	0.2215	0.2189	0.6200	0.6141	0.2439	0.2437
160	0.1651	0.1638	0.4445	0.4437	0.1797	0.1799
200	0.1299	0.1316	0.3476	0.3457	0.1410	0.1430
240	0.1092	0.1096	0.2866	0.2863	0.1178	0.1185
280	0.0934	0.0950	0.2444	0.2502	0.1005	0.1021
320	0.0817	0.0803	0.2112	0.2120	0.0878	0.0860
360	0.0730	0.0723	0.1884	0.1835	0.0779	0.0775
400	0.0656	0.0657	0.1685	0.1674	0.0700	0.0701
440	0.0605	0.0600	0.1517	0.1525	0.0645	0.0637
480	0.0552	0.0552	0.1380	0.1390	0.0589	0.0586

It is important to highlight that the average MSE of the lower limit and upper limit are approximately the same within each method, indicating same precision for the lower and upper limit estimates. It is also important to note that the MSE for all the three methods is monotonically decreasing as sample size increases. In addition, the MSE for the parametric method is asymptotically consistent, as shown below.

Let $x_1,\ldots,x_n$ be a random sample from a $N\left(\mu,\sigma^2\right)$ distribution, and let $\bar{x}=\frac{1}{n}\sum_{n}^{1}x_i$ and $s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$. Then the MSE of the lower limit estimates $\left(\hat{\theta}_L\right)$ produced by the parametric method can be derived as follows:

$$\mathrm{MSE}\left(\hat{\theta}_L\right)=\mathrm{Var}\left(\hat{\theta}_L\right)+\left(\mathrm{Bias}\left(\hat{\theta}_L-\theta_L\right)\right)^2$$

Note that

$$\begin{aligned}Var\left(\hat{\theta}_L\right)&=Var\left(\bar{x}-z_{\alpha/2}s\right)\\&=Var\left(\bar{x}\right)+z_{\alpha/2}^2Var\left(s\right)\end{aligned}$$

Recall that $\bar{x}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$, consequently

$$\begin{aligned}Var\left(\hat{\theta}_L\right)&=\frac{\sigma^2}{n}+z_{\alpha/2}^2Var\left(\frac{\sigma}{\sqrt{n-1}}\frac{\sqrt{n-1}}{\sigma}s\right)\\&=\frac{\sigma^2}{n}+z_{\alpha/2}^2\frac{\sigma^2}{n-1}Var\left(\frac{\sqrt{n-1}}{\sigma}s\right).\end{aligned}$$

Also, recall that $\frac{\sqrt{n-1}}{\sigma}s\sim\chi_{n-1}$. As a result,

$$Var\left(\hat{\theta}_L\right) = \frac{\sigma^2}{n} + z_{\alpha/2}^2 \frac{\sigma^2}{n-1}\left((n-1) - \mu^2\right)$$

$$= \frac{\sigma^2}{n} + z_{\alpha/2}^2 \frac{\sigma^2}{n-1}\left((n-1) - \left(\sqrt{2}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right)$$

$$= \frac{\sigma^2}{n} + z_{\alpha/2}^2 \frac{\sigma^2}{n-1}\left((n-1) - 2\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right).$$

Thus,

$$\mathrm{MSE}\left(\hat{\theta}_L\right) = \frac{\sigma^2}{n} + z_{\alpha/2}^2 \frac{\sigma^2}{n-1}\left((n-1) - 2\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right) + \left(z_{\alpha/2}\sigma\left(1 - \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)\right)^2$$

$$= \frac{\sigma^2}{n} + z_{\alpha/2}^2 \frac{\sigma^2}{n-1}\left((n-1) - 2\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right) + z_{\alpha/2}^2\sigma^2\left(1 - \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2$$

$$= \sigma^2\left(\frac{1}{n} + z_{\alpha/2}^2 \frac{1}{n-1}\left((n-1) - 2\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right) + z_{\alpha/2}^2\left(1 - \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right) \quad (4)$$

Similarly, it can be shown that the MSE of the upper limit is equal to the MSE of the lower limit. Now recall (3):

$$\lim_{n\to\infty}\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}=1.$$

This indicates that the MSE of both the lower and upper limit estimates converge to 0. Thus, the estimates of the lower and upper limits produced by the parametric method are indeed asymptotically consistent.

Note that the results presented in Tables 3 and 4 and Figures 4.1 to 4.7 are, without loss of generality, for data from a $N(20,9)$ distribution. The findings are similar for different means and variances in the sense that the parametric method performed better by looking at both bias and MSE. The general pattern of bias, that is, bias decreasing with sample size and the asymptotic unbiasedness of the parametric approach still holds for different means and variances. The results also indicate that the non-parametric and robust methods are asymptotically biased (although the bias is very small) regardless of mean and variance. In addition, the MSE of the three methods monotonically decreases with sample size, regardless of mean and variance, with the parametric method performing uniformly better followed by the robust approach. Moreover, the parametric method is asymptotically consistent for different means and variances.

Nevertheless, we observed that the magnitudes of both bias and MSE increase with variance, but both are not affected by the change in mean. Further investigation confirmed that the increases in both bias and MSE are directly proportional to variance and inversely proportional to sample size. We were able to empirically show that the

standardized bias and MSE (Bias/σ and MSE/σ^2) are constant for all three methods for a given sample size. In fact, we were able to show this analytically for the parametric method. Recall the bias of the estimated lower limit (1):

$$\text{Bias}\left(\hat{\theta}_L, \theta_L\right) = z_{\alpha/2}\sigma\left(1 - \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right).$$

Now, if we standardize (divide) the bias of the estimated lower limit by dividing by σ, we end up with the expression

$$\frac{\text{Bias}\left(\hat{\theta}_L, \theta_L\right)}{\sigma} = z_{\alpha/2}\left(1 - \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right). \quad (5)$$

Note that (5) is a function of sample size only. Similarly, standardizing the bias of the estimated upper limit (2) by dividing by σ returns a function of sample size only:

$$\frac{\text{Bias}\left(\hat{\theta}_U, \theta_U\right)}{\sigma} = z_{\alpha/2}\left(\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} - 1\right). \quad (6)$$

Note that the right hand side of (5) and (6), respectively, are bias of the lower and upper bound estimates, if a standard normal distribution was used.

Recall the MSE of the estimated lower limit (4):

$$\mathrm{MSE}\left(\hat{\theta}_L\right) = \sigma^2 \left(\frac{1}{n} + z_{\alpha/2}^2 \frac{1}{n-1} \left((n-1) - 2 \left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \right)^2 \right) + z_{\alpha/2}^2 \left(1 - \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \right)^2 \right).$$

Now, if we standardize the MSE of the estimated lower limit estimates by dividing with σ^2, we end up with the expression

$$\frac{\mathrm{MSE}\left(\hat{\theta}_L\right)}{\sigma^2} = \frac{1}{n} + z_{\alpha/2}^2 \frac{1}{n-1} \left((n-1) - 2 \left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \right)^2 \right) + z_{\alpha/2}^2 \left(1 - \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \right)^2 . \quad (7)$$

Similarly, it can be shown that the standardized MSE of the upper limit is equal to the standardized MSE of the lower limit. Note that (7) is a function of sample size only and is the MSE that will be obtained if data from a standard normal distribution is used. Table 5 provides the average bias and MSE of the non-parametric method for each of the three Gaussian distributions considered and Table 6 provides the standardized average bias and MSE. The standardized average bias for the estimated lower limit is approximately the same regardless of the distribution, as it is for the estimated upper limit. The same is true for the standardized average MSE.

Table 5: *Average empirical bias and MSE for non-parametric method, where data is generated from* $N(2.45,0.01)$, $N(20,9)$, $N(74.82,132.70)$.

n	Average Bias						Average MSE					
	$N(2.45,0.01)$		$N(20,9)$		$N(74.82,132.70)$		$N(2.45,0.01)$		$N(20,9)$		$N(74.82,132.70)$	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	-0.0181	0.0178	-0.6042	0.5940	-2.3200	2.2810	0.0022	0.0021	2.4101	2.3821	35.5343	35.1213
80	-0.0088	0.0089	-0.2927	0.2971	-1.1240	1.1407	0.0009	0.0009	1.0050	0.9941	14.8181	14.6574
120	-0.0056	0.0057	-0.1879	0.1903	-0.7216	0.7308	0.0006	0.0006	0.6200	0.6141	9.1411	9.0542
160	-0.0042	0.0041	-0.1410	0.1380	-0.5415	0.5298	0.0004	0.0004	0.4445	0.4437	6.5539	6.5425
200	-0.0034	0.0033	-0.1148	0.1115	-0.4407	0.4283	0.0003	0.0003	0.3476	0.3457	5.1250	5.0977
240	-0.0028	0.0028	-0.0939	0.0948	-0.3607	0.3639	0.0003	0.0003	0.2866	0.2863	4.2261	4.2209
280	-0.0025	0.0025	-0.0836	0.0827	-0.3212	0.3176	0.0002	0.0002	0.2444	0.2502	3.6030	3.6892
320	-0.0020	0.0022	-0.0677	0.0729	-0.2598	0.2798	0.0002	0.0002	0.2112	0.2120	3.1142	3.1262
360	-0.0019	0.0019	-0.0647	0.0617	-0.2485	0.2370	0.0002	0.0002	0.1884	0.1835	2.7780	2.7061
400	-0.0017	0.0017	-0.0566	0.0571	-0.2175	0.2194	0.0002	0.0002	0.1685	0.1674	2.4838	2.4686
440	-0.0014	0.0016	-0.0481	0.0531	-0.1849	0.2041	0.0001	0.0001	0.1517	0.1525	2.2366	2.2489
480	-0.0014	0.0014	-0.0477	0.0477	-0.1833	0.1832	0.0001	0.0001	0.1380	0.1390	2.0351	2.0493

Table 6: *Standardized average empirical bias and MSE for non-parametric method, where data is generated from* $N(2.45,0.01)$, $N(20,9)$, $N(74.82,132.70)$.

n	Average Bias/σ						Average MSE/σ^2					
	$N(2.45,0.01)$		$N(20,9)$		$N(74.82,132.70)$		$N(2.45,0.01)$		$N(20,9)$		$N(74.82,132.70)$	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	-0.2014	0.1980	-0.2014	0.1980	-0.2014	0.1980	0.2678	0.2647	0.2678	0.2647	0.2678	0.2647
80	-0.0976	0.0990	-0.0976	0.0990	-0.0976	0.0990	0.1117	0.1105	0.1117	0.1105	0.1117	0.1105
120	-0.0626	0.0634	-0.0626	0.0634	-0.0626	0.0634	0.0689	0.0682	0.0689	0.0682	0.0689	0.0682
160	-0.0470	0.0460	-0.0470	0.0460	-0.0470	0.0460	0.0494	0.0493	0.0494	0.0493	0.0494	0.0493
200	-0.0383	0.0372	-0.0383	0.0372	-0.0383	0.0372	0.0386	0.0384	0.0386	0.0384	0.0386	0.0384
240	-0.0313	0.0316	-0.0313	0.0316	-0.0313	0.0316	0.0318	0.0318	0.0318	0.0318	0.0318	0.0318
280	-0.0279	0.0276	-0.0279	0.0276	-0.0279	0.0276	0.0272	0.0278	0.0272	0.0278	0.0272	0.0278
320	-0.0225	0.0243	-0.0226	0.0243	-0.0226	0.0243	0.0235	0.0236	0.0235	0.0236	0.0235	0.0236
360	-0.0216	0.0206	-0.0216	0.0206	-0.0216	0.0206	0.0209	0.0204	0.0209	0.0204	0.0209	0.0204
400	-0.0189	0.0190	-0.0189	0.0190	-0.0189	0.0190	0.0187	0.0186	0.0187	0.0186	0.0187	0.0186
440	-0.0160	0.0177	-0.0160	0.0177	-0.0160	0.0177	0.0169	0.0169	0.0169	0.0169	0.0169	0.0169
480	-0.0159	0.0159	-0.0159	0.0159	-0.0159	0.0159	0.0153	0.0154	0.0153	0.0154	0.0153	0.0154

Tables 7 and 8 provide the average bias and MSE and standardized average bias and MSE, respectively, for the robust method for each of the three Gaussian distributions considered. Similar to the non-parametric method, the standardized average bias for the estimated lower limit is approximately the same regardless of the distribution, as it is for the estimated upper limit. The same is true for the standardized average MSE.

***Table 7**: Average empirical bias and MSE for robust method, where data is generated from* $N(2.45,0.01)$, $N(20,9)$, $N(74.82,132.70)$.

n	Average Bias						Average MSE					
	$N(2.45,0.01)$		$N(20,9)$		$N(74.82,132.70)$		$N(2.45,0.01)$		$N(20,9)$		$N(74.82,132.70)$	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	-0.0080	0.0081	-0.2677	0.2715	-1.0278	1.0427	0.0007	0.0007	0.8218	0.8259	12.1170	12.1778
80	-0.0039	0.0041	-0.1287	0.1351	-0.4940	0.5189	0.0003	0.0003	0.3717	0.3777	5.4800	5.5686
120	-0.0025	0.0027	-0.0828	0.0904	-0.3179	0.3470	0.0002	0.0002	0.2439	0.2437	3.5956	3.5938
160	-0.0019	0.0020	-0.0628	0.0660	-0.2413	0.2533	0.0002	0.0002	0.1797	0.1799	2.6501	2.6525
200	-0.0016	0.0016	-0.0540	0.0535	-0.2074	0.2053	0.0001	0.0001	0.1410	0.1430	2.0784	2.1080
240	-0.0013	0.0014	-0.0431	0.0470	-0.1654	0.1805	0.0001	0.0001	0.1178	0.1185	1.7365	1.7476
280	-0.0011	0.0012	-0.0381	0.0401	-0.1463	0.1541	0.0001	0.0001	0.1005	0.1021	1.4817	1.5060
320	-0.0010	0.0010	-0.0330	0.0339	-0.1267	0.1301	0.0001	0.0001	0.0878	0.0860	1.2939	1.2687
360	-0.0009	0.0009	-0.0292	0.0304	-0.1122	0.1169	0.0001	0.0001	0.0779	0.0775	1.1490	1.1427
400	-0.0008	0.0008	-0.0260	0.0268	-0.1000	0.1029	0.0001	0.0001	0.0700	0.0701	1.0325	1.0335
440	-0.0007	0.0007	-0.0243	0.0241	-0.0934	0.0924	0.0001	0.0001	0.0645	0.0637	0.9506	0.9395
480	-0.0007	0.0007	-0.0226	0.0219	-0.0868	0.0843	0.0001	0.0001	0.0589	0.0586	0.8687	0.8643

Table 8: *Standardized average empirical bias and MSE for robust method, where data is generated from* $N(2.45,0.01)$, $N(20,9)$, $N(74.82,132.70)$.

n	Average Bias/σ						Average MSE/σ^2					
	$N(2.45,0.01)$		$N(20,9)$		$N(74.82,132.70)$		$N(2.45,0.01)$		$N(20,9)$		$N(74.82,132.70)$	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	-0.0892	0.0905	-0.0892	0.0905	-0.0892	0.0905	0.0913	0.0918	0.0913	0.0918	0.0913	0.0918
80	-0.0429	0.0450	-0.0429	0.0450	-0.0429	0.0450	0.0413	0.0420	0.0413	0.0420	0.0413	0.0420
120	-0.0276	0.0301	-0.0276	0.0301	-0.0276	0.0301	0.0271	0.0271	0.0271	0.0271	0.0271	0.0271
160	-0.0209	0.0220	-0.0209	0.0220	-0.0209	0.0220	0.0200	0.0200	0.0200	0.0200	0.0200	0.0200
200	-0.0180	0.0178	-0.0180	0.0178	-0.0180	0.0178	0.0157	0.0159	0.0157	0.0159	0.0157	0.0159
240	-0.0144	0.0157	-0.0144	0.0157	-0.0144	0.0157	0.0131	0.0132	0.0131	0.0132	0.0131	0.0132
280	-0.0127	0.0134	-0.0127	0.0134	-0.0127	0.0134	0.0112	0.0114	0.0112	0.0113	0.0112	0.0113
320	-0.0110	0.0113	-0.0110	0.0113	-0.0110	0.0113	0.0098	0.0096	0.0098	0.0096	0.0098	0.0096
360	-0.0097	0.0101	-0.0097	0.0101	-0.0097	0.0101	0.0087	0.0086	0.0087	0.0086	0.0087	0.0086
400	-0.0087	0.0089	-0.0087	0.0089	-0.0087	0.0089	0.0078	0.0078	0.0078	0.0078	0.0078	0.0078
440	-0.0081	0.0080	-0.0081	0.0080	-0.0081	0.0080	0.0072	0.0071	0.0072	0.0071	0.0072	0.0071
480	-0.0075	0.0073	-0.0075	0.0073	-0.0075	0.0073	0.0065	0.0065	0.0065	0.0065	0.0065	0.0065

Additional simulations were conducted, where data is generated from a standard normal distribution. The standardized average bias and MSE for all three methods are in fact consistent with the average bias and MSE produced by the standardized normal distribution (Table 9). Consequently, without loss of generality, results regarding the performances of the methods from the standard normal distribution can be extended to Gaussian distributions with any mean and variance.

***Table 9**: Average empirical bias and MSE for all three methods, where data is generated from $N(0,1)$.*

n	Average Bias						Average MSE					
	Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	0.0133	-0.0128	-0.2014	0.1980	-0.0892	0.0905	0.0740	0.0739	0.2678	0.2647	0.0913	0.0918
80	0.0076	-0.0060	-0.0976	0.0990	-0.0429	0.0450	0.0364	0.0365	0.1117	0.1105	0.0413	0.0420
120	0.0058	-0.0039	-0.0626	0.0634	-0.0276	0.0301	0.0246	0.0243	0.0689	0.0682	0.0271	0.0271
160	0.0040	-0.0036	-0.0470	0.0460	-0.0210	0.0220	0.0183	0.0182	0.0494	0.0493	0.0200	0.0200
200	0.0020	-0.0025	-0.0383	0.0372	-0.0180	0.0178	0.0144	0.0146	0.0386	0.0384	0.0157	0.0159
240	0.0024	-0.0012	-0.0313	0.0316	-0.0144	0.0157	0.0121	0.0122	0.0318	0.0318	0.0131	0.0132
280	0.0016	-0.0011	-0.0279	0.0276	-0.0127	0.0134	0.0104	0.0106	0.0272	0.0278	0.0112	0.0113
320	0.0015	-0.0014	-0.0226	0.0243	-0.0110	0.0113	0.0091	0.0089	0.0235	0.0236	0.0098	0.0096
360	0.0014	-0.0011	-0.0216	0.0206	-0.0097	0.0101	0.0081	0.0080	0.0209	0.0204	0.0087	0.0086
400	0.0013	-0.0012	-0.0189	0.0190	-0.0087	0.0089	0.0073	0.0073	0.0187	0.0186	0.0078	0.0078
440	0.0012	-0.0010	-0.0160	0.0177	-0.0081	0.0080	0.0067	0.0067	0.0169	0.0169	0.0072	0.0071
480	0.0008	-0.0011	-0.0159	0.0159	-0.0075	0.0073	0.0061	0.0061	0.0153	0.0154	0.0065	0.0065

In addition, it is important to note that the standardized bias and MSE for the lower and upper limits, obtained theoretically (analytically) (equations 3,4, and 6) (results presented in Table 10) are close to the average bias and MSE produced empirically by the parametric method for a standard normal data (Table 9), indicating that the empirical bias and MSE are good estimates of the true bias and MSE.

Table 10: *Theoretical standardized bias and MSE for parametric method.*

n	Bias		MSE	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	0.0125	-0.0125	0.0741	0.0741
80	0.0062	-0.0062	0.0368	0.0368
120	0.0041	-0.0041	0.0245	0.0245
160	0.0031	-0.0031	0.0183	0.0183
200	0.0025	-0.0025	0.0146	0.0146
240	0.0020	-0.0020	0.0122	0.0122
280	0.0018	-0.0018	0.0105	0.0105
320	0.0015	-0.0015	0.0091	0.0091
360	0.0014	-0.0014	0.0081	0.0081
400	0.0012	-0.0012	0.0073	0.0073
440	0.0011	-0.0011	0.0066	0.0066
480	0.0010	-0.0010	0.0061	0.0061

Finally, consider the coverage probability and width of the 90% confidence intervals produced by the parametric, non-parametric, and robust methods. Coverage probabilities for confidence intervals of the lower and upper limits estimated by all three methods are provided in Table 11.

Table 11: *Coverage probability for all three methods, where data is generated from N(20,9).*

n	Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	0.89	0.91	N/A	N/A	0.88	0.88
80	0.90	0.90	N/A	N/A	0.88	0.89
120	0.92	0.91	0.92	0.92	0.90	0.89
160	0.89	0.90	0.92	0.93	0.88	0.89
200	0.89	0.90	0.94	0.93	0.88	0.89
240	0.91	0.91	0.95	0.96	0.90	0.90
280	0.90	0.91	0.95	0.94	0.90	0.89
320	0.90	0.90	0.92	0.93	0.89	0.89
360	0.89	0.91	0.95	0.94	0.88	0.89
400	0.88	0.89	0.94	0.94	0.87	0.89
440	0.91	0.91	0.94	0.94	0.90	0.90
480	0.91	0.91	0.92	0.91	0.89	0.89
avg	0.90	0.90	0.94	0.93	0.89	0.89

Note that because 120 samples are required to provide confidence intervals for limits estimated by the non-parametric method, coverage probabilities could not be provided when $n = 40$ and $n = 80$. For all three methods, there is no evident relationship between sample size and coverage probabilities. The average (avg) coverage probabilities for all three methods are very close to the nominal coverage probability of 0.90. However, the

coverage probabilities of the non-parametric confidence intervals are consistently larger than the nominal coverage probability. This is most likely because the widths of the confidence intervals produced by the non-parametric method are consistently wider than the widths of the confidence intervals produced by the parametric and robust methods, as shown in Table 12.

***Table 12**: Confidence interval widths for all three methods, where data is generated from N(20,9).*

n	Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	2.66	2.66	N/A	N/A	2.79	2.80
80	1.89	1.89	N/A	N/A	1.95	1.96
120	1.54	1.54	2.90	2.92	1.59	1.59
160	1.34	1.34	2.98	2.98	1.37	1.37
200	1.19	1.19	2.24	2.25	1.22	1.22
240	1.09	1.09	2.29	2.40	1.11	1.11
280	1.01	1.01	1.98	1.97	1.03	1.03
320	0.94	0.94	1.66	1.67	0.97	0.97
360	0.89	0.89	1.73	1.77	0.91	0.90
400	0.84	0.84	1.53	1.53	0.86	0.86
440	0.80	0.80	1.46	1.47	0.82	0.82
480	0.77	0.77	1.33	1.33	0.78	0.79

The widths of the confidence intervals produced by all three methods have an inversely proportional relationship with sample size. The parametric method uniformly provides narrower confidence intervals than the non-parametric and robust methods. In addition, the widths of the confidence intervals of the lower and upper limit produced by the parametric method are exactly the same for each sample size. Confidence intervals that accompany the robust method are slightly wider than the confidence intervals that accompany the parametric method. As sample size increases, the difference between the widths of these two methods becomes negligible. The non-parametric method, on the other hand, produces confidence intervals that are significantly wider than the confidence intervals provided by the parametric and robust methods.

4.4.2 Results for Skew Normal Distributions

A total of fifteen skew normal distributions were considered using combinations of different mean, variance and skewness parameters. The means and variances considered are the same as those in the Gaussian scenarios, and 5 levels of skewness (κ) were integrated into the simulation: $\kappa = 0.10, 0.25, 0.50, 0.75, 0.95$. Similar to the simulation for the Gaussian scenarios, 1000 datasets were generated from each of the skewed distributions with sample sizes of $n = 40, 80, 120, 160, 200, 240, 280, 320, 360, 400, 440,$ and 480. RIs were estimated for each dataset using the parametric, non-parametric, and robust methods. The bias and MSE of the lower and upper limit were computed and subsequently averaged over additional 50 samples (repetitions).

First consider data from the skew normal distribution, where skewness is very small ($\kappa = 0.1$). Plots of bias for lower and upper limits estimated by the parametric, non-parametric and robust methods are provided in Figure 4.8. As can be seen from this figure, bias in general remains small for all three methods when data is slightly skewed ($\kappa = 0.10$). However, compared to the bias of estimates from Gaussian data, the magnitude of bias has increased. Overall, the parametric method appears to perform better than the non-parametric approach when estimating both the lower and upper limits with small sample size. However, with large sample sizes ($n \geq 200$), the non-parametric approach provides the smallest bias for estimates of both limits (Table 13). The robust method does not seem to perform well in estimating the lower limit compared to both the parametric and non-parametric approaches, regardless of sample size. However, the robust method provides the smallest bias for estimates of the upper limit with small sample sizes ($n < 200$).

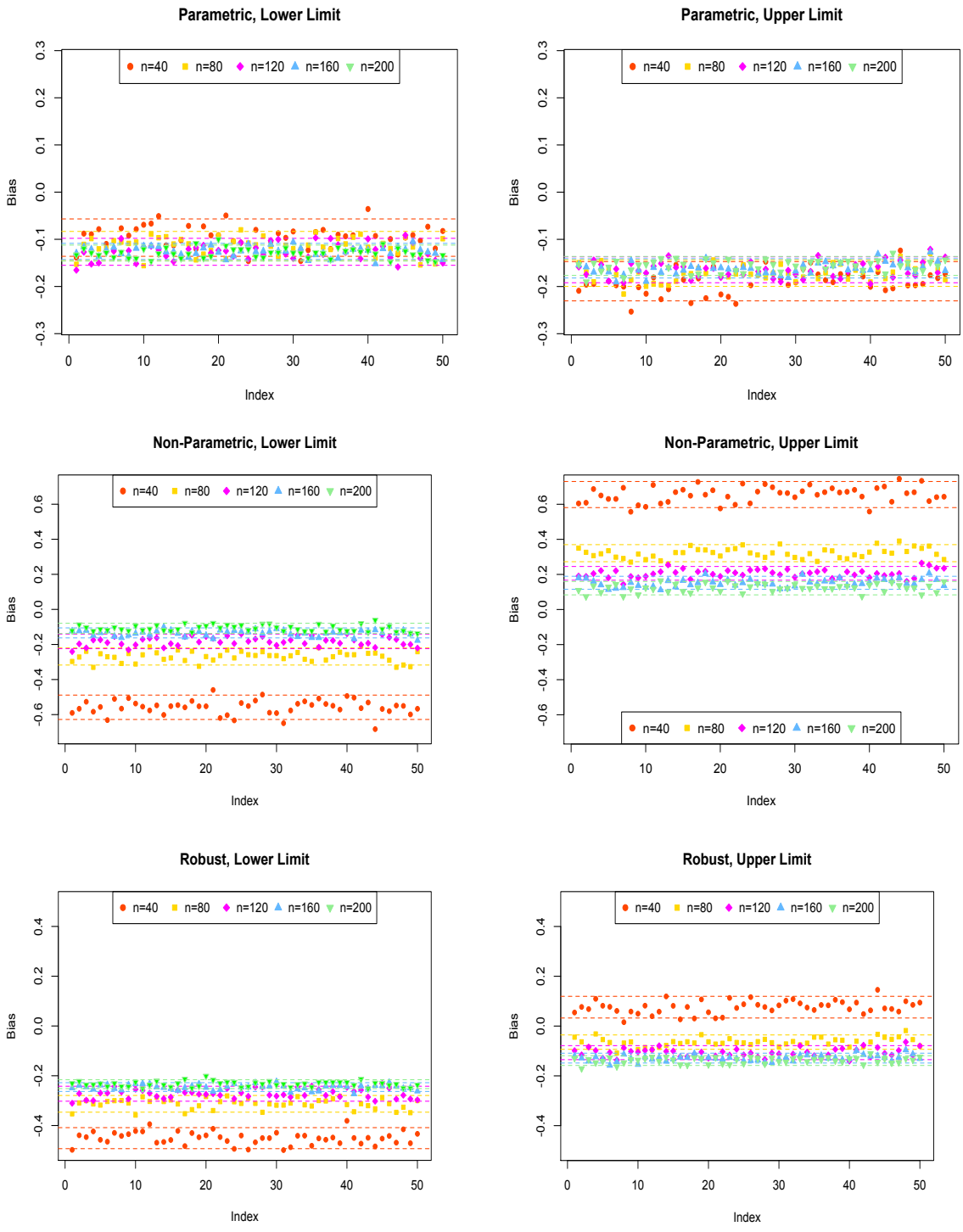

Figure 4.8: *Empirical bias, with corresponding 95% confidence intervals (indicated by dashed lines), for the three methods, where data is generated from skew normal distribution with κ = 0.1. The mean and variance are 20 and 9, respectively.*

For the parametric method, a change in direction of bias associated with the estimates of the lower limit was observed. Instead of the negligible, positive bias observed in the Gaussian scenarios, small, negative bias is observed with slightly positive skewed data. The upper limit estimates continue to have negative bias with skewed data, as did the estimates with Gaussian data. A change in direction of bias is also observed for the robust method, except the change is associated with the estimates of the upper limit. Negative bias is observed for skewed data (excluding $n = 40$), contrary to the positive bias observed with Gaussian data. The lower limit estimates continue to have negative bias with skewed data, as did the estimates with Gaussian data. Due to this shift in the direction of bias, both the parametric and robust methods lead to false negatives (missed treatment) near the lower limit and false positives (unnecessary treatment) near the upper limit, when data is skewed. The non-parametric method, on the other hand, maintains the same the direction of bias for both skewed and Gaussian distributions.

Similar to the results observed with Gaussian data, the variability of bias decreased with an increase in sample size (Figure 4.8) and bias appeared to converge to a constant for all three methods. However, the constant that bias converges to appears to be a non-zero constant, indicating that all three methods are asymptotically biased. This is contrary to the Gaussian scenarios, where the parametric method resulted in asymptotically unbiased estimators.

Similar patterns in bias are observed as the skewness parameter increases ($\kappa = 0.25$). However, the magnitude of bias of the estimated lower and upper limits has increased for

all three methods, except for lower limits estimated by the non-parametric method (Table 13). Note that the robust method now estimates the upper limit with negative bias on average when $n = 40$. In addition, the average bias associated with the non-parametric method appears to be smaller than the parametric and robust methods, even for a sample size as small as $n = 80$.

***Table 13**: Average empirical bias for the three different methods, where data is generated from skew normal distributions with small levels of skewness. The mean and variance are 20 and 9, respectively.*

n	Skew = 0.1						Skew = 0.25					
	Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	-0.0964	-0.1888	-0.5583	0.6553	-0.4507	0.0759	-0.3039	-0.4171	-0.5034	0.7271	-0.7289	-0.2176
80	-0.1141	-0.1715	-0.2684	0.3212	-0.3128	-0.0644	-0.3297	-0.3972	-0.2453	0.3549	-0.5996	-0.3538
120	-0.1265	-0.1645	-0.1823	0.2062	-0.2722	-0.1068	-0.3398	-0.3897	-0.1659	0.2282	-0.5570	-0.3954
160	-0.1257	-0.1615	-0.1340	0.1524	-0.2455	-0.1287	-0.3376	-0.3876	-0.1205	0.1680	-0.5294	-0.4185
200	-0.1291	-0.1564	-0.1081	0.1217	-0.2337	-0.1386	-0.3397	-0.3825	-0.0946	0.1351	-0.5160	-0.4282
240	-0.1313	-0.1548	-0.0887	0.1012	-0.2259	-0.1472	-0.3439	-0.3801	-0.0818	0.1125	-0.5101	-0.4356
280	-0.1318	-0.1532	-0.0747	0.0931	-0.2205	-0.1539	-0.3445	-0.3785	-0.0661	0.1031	-0.5051	-0.4422
320	-0.1338	-0.1540	-0.0683	0.0775	-0.2160	-0.1592	-0.3449	-0.3794	-0.0631	0.0863	-0.4990	-0.4477
360	-0.1308	-0.1535	-0.0575	0.0672	-0.2090	-0.1632	-0.3433	-0.3791	-0.0522	0.0740	-0.4935	-0.4516
400	-0.1364	-0.1532	-0.0567	0.0611	-0.2112	-0.1661	-0.3477	-0.3783	-0.0518	0.0682	-0.4945	-0.4543
440	-0.1323	-0.1529	-0.0459	0.0536	-0.2042	-0.1685	-0.3445	-0.3785	-0.0418	0.0592	-0.4884	-0.4571
480	-0.1325	-0.1526	-0.0429	0.0516	-0.2020	-0.1704	-0.3446	-0.3782	-0.0363	0.0571	-0.4864	-0.4592

Now, consider moderately skewed data ($\kappa = 0.50$). The magnitude of bias has increased compared to symmetric data as well as data with mild skewness ($\kappa = 0.1$ and $\kappa = 0.25$). In fact, we considered even higher skewness levels and observed that the magnitude of bias generally increases with skewness (Table 14), with the exception of the non-parametric method, where bias associated with the lower limit decreases when skewness increases. This, as expected, is due to the positive skewness in the data, where many of the data points are accumulated in the lower level percentiles leading to more sample size for the lower limit estimates. The direction of bias remained the same for all levels of skewness. We would also like to highlight that the variability of bias decreases with sample size, regardless of the magnitude of skewness in the data, and bias converges to non-zero constants for all three methods, indicating asymptotically biased estimates. The bias for the non-parametric method converges to a magnitude smaller than both the parametric and non-parametric methods indicating that when data are skewed, it asymptotically produces better estimates of lower and upper limits.

Table 14: *Average empirical bias for the three different methods, where data is generated from skew normal distributions with moderate to large levels of skewness. The mean and variance are 20 and 9, respectively.*

n	Skew = 0.50						Skew = 0.75						Skew = 0.95					
	Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	-0.7015	-0.7697	-0.4196	0.8106	-1.2595	-0.6735	-1.1969	-1.0824	-0.3023	0.8742	-1.9117	-1.0753	-1.7957	-1.3096	-0.1409	0.9173	-2.6453	-1.3527
80	-0.7256	-0.7485	-0.1981	0.3958	-1.1310	-0.8057	-1.2239	-1.0591	-0.1479	0.4268	-1.7877	-1.2009	-1.8261	-1.2836	-0.0707	0.4478	-2.5214	-1.4687
120	-0.7323	-0.7415	-0.1323	0.2546	-1.0850	-0.8467	-1.2306	-1.0517	-0.0994	0.2746	-1.7420	-1.2398	-1.8321	-1.2762	-0.0482	0.2881	-2.4737	-1.5051
160	-0.7360	-0.7378	-0.1028	0.1873	-1.0630	-0.8670	-1.2323	-1.0485	-0.0740	0.2020	-1.7180	-1.2600	-1.8345	-1.2726	-0.0343	0.2120	-2.4498	-1.5241
200	-0.7370	-0.7325	-0.0797	0.1505	-1.0488	-0.8766	-1.2344	-1.0426	-0.0575	0.1624	-1.7049	-1.2684	-1.8371	-1.2662	-0.0274	0.1704	-2.4367	-1.5312
240	-0.7389	-0.7306	-0.0662	0.1253	-1.0407	-0.8842	-1.2367	-1.0404	-0.0507	0.1352	-1.6970	-1.2751	-1.8380	-1.2645	-0.0225	0.1418	-2.4276	-1.5381
280	-0.7420	-0.7281	-0.0554	0.1148	-1.0384	-0.8898	-1.2413	-1.0370	-0.0430	0.1238	-1.6963	-1.2796	-1.8435	-1.2606	-0.0200	0.1300	-2.4275	-1.5417
320	-0.7397	-0.7301	-0.0478	0.0960	-1.0299	-0.8966	-1.2375	-1.0399	-0.0357	0.1036	-1.6866	-1.2870	-1.8407	-1.2631	-0.0164	0.1087	-2.4184	-1.5485
360	-0.7400	-0.7293	-0.0405	0.0824	-1.0263	-0.8997	-1.2382	-1.0388	-0.0292	0.0889	-1.6835	-1.2897	-1.8423	-1.2616	-0.0144	0.0933	-2.4159	-1.5505
400	-0.7429	-0.7286	-0.0393	0.0760	-1.0256	-0.9025	-1.2409	-1.0381	-0.0302	0.0819	-1.6828	-1.2926	-1.8436	-1.2615	-0.0141	0.0860	-2.4140	-1.5540
440	-0.7419	-0.7283	-0.0360	0.0659	-1.0216	-0.9043	-1.2401	-1.0378	-0.0260	0.0711	-1.6788	-1.2940	-1.8429	-1.2611	-0.0118	0.0746	-2.4101	-1.5552
480	-0.7419	-0.7281	-0.0321	0.0637	-1.0196	-0.9066	-1.2402	-1.0374	-0.0239	0.0687	-1.6769	-1.2961	-1.8436	-1.2604	-0.0110	0.0720	-2.4086	-1.5569

Now consider the MSE when skewness is small ($\kappa = 0.1$). When estimating the lower limit, the parametric method uniformly has minimum MSE, compared to the non-parametric and robust methods. When estimating the upper limit, the difference in MSE between the parametric and robust methods is negligible (Figure 4.9, Table 15). The non-parametric method uniformly has the largest MSE when estimating both the lower and upper limits. This shows that, although the non-parametric method produces less biased estimates for larger sample sizes, it also produces less precise estimates when data is slightly skewed. Careful consideration of the three methods is therefore important.

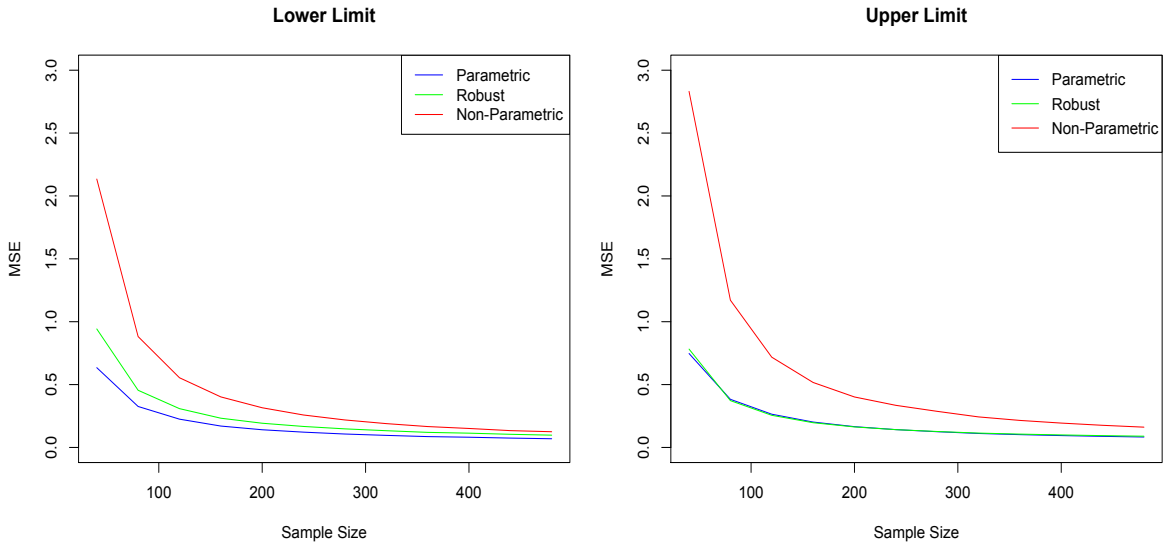

Figure 4.9: *Average empirical MSE for the three methods, where data is generated from skewed normal distribution with $\kappa = 0.1$. The mean and variance are 20 and 9, respectively.*

Table 15: *Average empirical MSE for the three different methods, where data is generated from skew normal distributions with small levels of skewness. The mean and variance are 20 and 9, respectively.*

n	Skew = 0.10						Skew = 0.25					
	Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	0.6345	0.7471	2.1344	2.8314	0.9424	0.7821	0.6754	0.9688	1.7673	3.4319	1.2755	0.8709
80	0.3258	0.3828	0.8817	1.1710	0.4546	0.3725	0.3978	0.5537	0.7380	1.4120	0.7147	0.5187
120	0.2250	0.2643	0.5543	0.7179	0.3086	0.2557	0.3085	0.4184	0.4668	0.8658	0.5434	0.4191
160	0.1702	0.2021	0.4019	0.5168	0.2326	0.1967	0.2584	0.3475	0.3348	0.6224	0.4542	0.3679
200	0.1407	0.1653	0.3157	0.4015	0.1926	0.1633	0.2319	0.3042	0.2646	0.4828	0.4053	0.3376
240	0.1217	0.1418	0.2580	0.3350	0.1670	0.1416	0.2155	0.2767	0.2168	0.4029	0.3764	0.3182
280	0.1068	0.1252	0.2187	0.2872	0.1478	0.1268	0.2021	0.2577	0.1836	0.3455	0.3550	0.3063
320	0.0960	0.1112	0.1896	0.2429	0.1332	0.1142	0.1918	0.2425	0.1606	0.2922	0.3359	0.2961
360	0.0863	0.1020	0.1662	0.2150	0.1199	0.1061	0.1816	0.2319	0.1414	0.2581	0.3191	0.2894
400	0.0813	0.0945	0.1510	0.1938	0.1136	0.0996	0.1791	0.2228	0.1298	0.2331	0.3137	0.2835
440	0.0745	0.0883	0.1338	0.1761	0.1046	0.0939	0.1715	0.2162	0.1150	0.2120	0.3014	0.2792
480	0.0693	0.0830	0.1251	0.1614	0.0978	0.0895	0.1669	0.2099	0.1033	0.1939	0.2939	0.2754

When skewness increases to $\kappa = 0.25$, no method is uniformly best across all sample sizes (Figure 4.10, Table 15). In terms of the lower limit, the parametric method is uniformly the best until a sample size of approximately $n = 240$ is reached, then the non-parametric method becomes uniformly the best afterwards. The robust method performs better than the non-parametric method up to a sample size of 80, and performs worst afterwards. When estimating the upper limit, the robust method performs the best for

small sample sizes (up to 120), where the difference in MSE between the robust and parametric is negligible. The parametric method performs the best for moderate to large sample sizes ($120 \leq n < 440$), with negligible to small difference in MSE with the robust. The parametric and non-parametric methods have comparable performances for very large sample sizes ($n \geq 440$), where the non-parametric provides a slightly less MSE (Table 15).

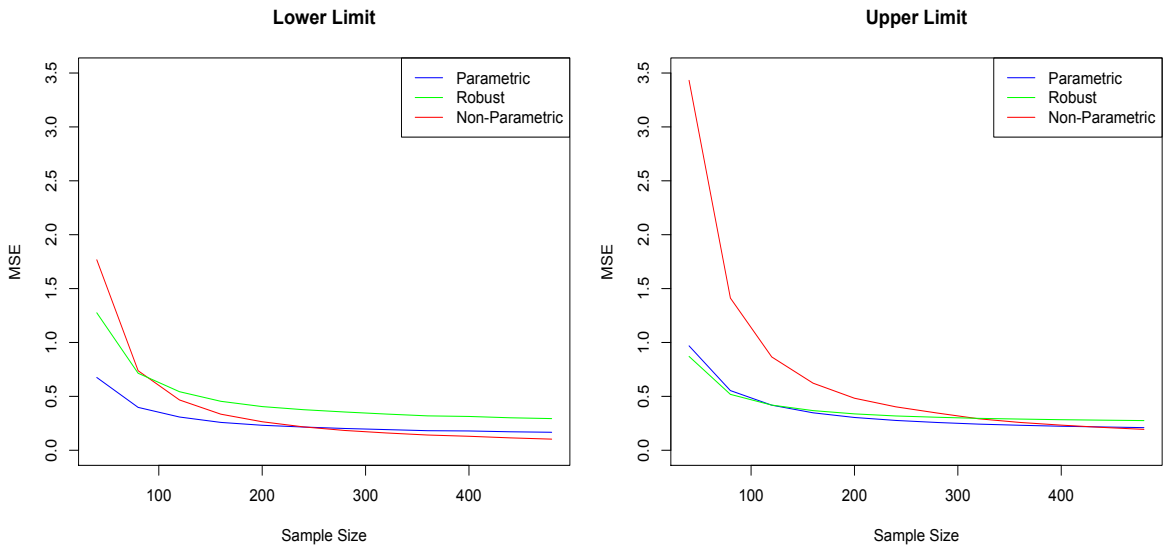

Figure 4.10: *Average empirical MSE for the three methods, where data is generated from skewed normal distribution with $\kappa = 0.25$. The mean and the variance are 20 and 9, respectively.*

As the skewness level in the data increases, the parametric method becomes sensitive to deviation from normality (symmetry) and its performance declines, and the non-parametric method starts to perform the best for most of the sample size considered. However, what is interesting to note from these results is that, the non-parametric method

does not have best performance at sample size of $n = 120$ (the sample size recommended by the Clinical Laboratory Standards Institute (CLSI)) for either of the RI limits. In fact, this is not observed until skewness is high ($\kappa = 0.75$) (Figure 4.11, Table 16). Thus, when estimating RIs with the non-parametric method for symmetric data as well as data with small to moderately skewed (skewness of less than 0.75), laboratories are under the mistaken impression that they are producing the best estimates possible.

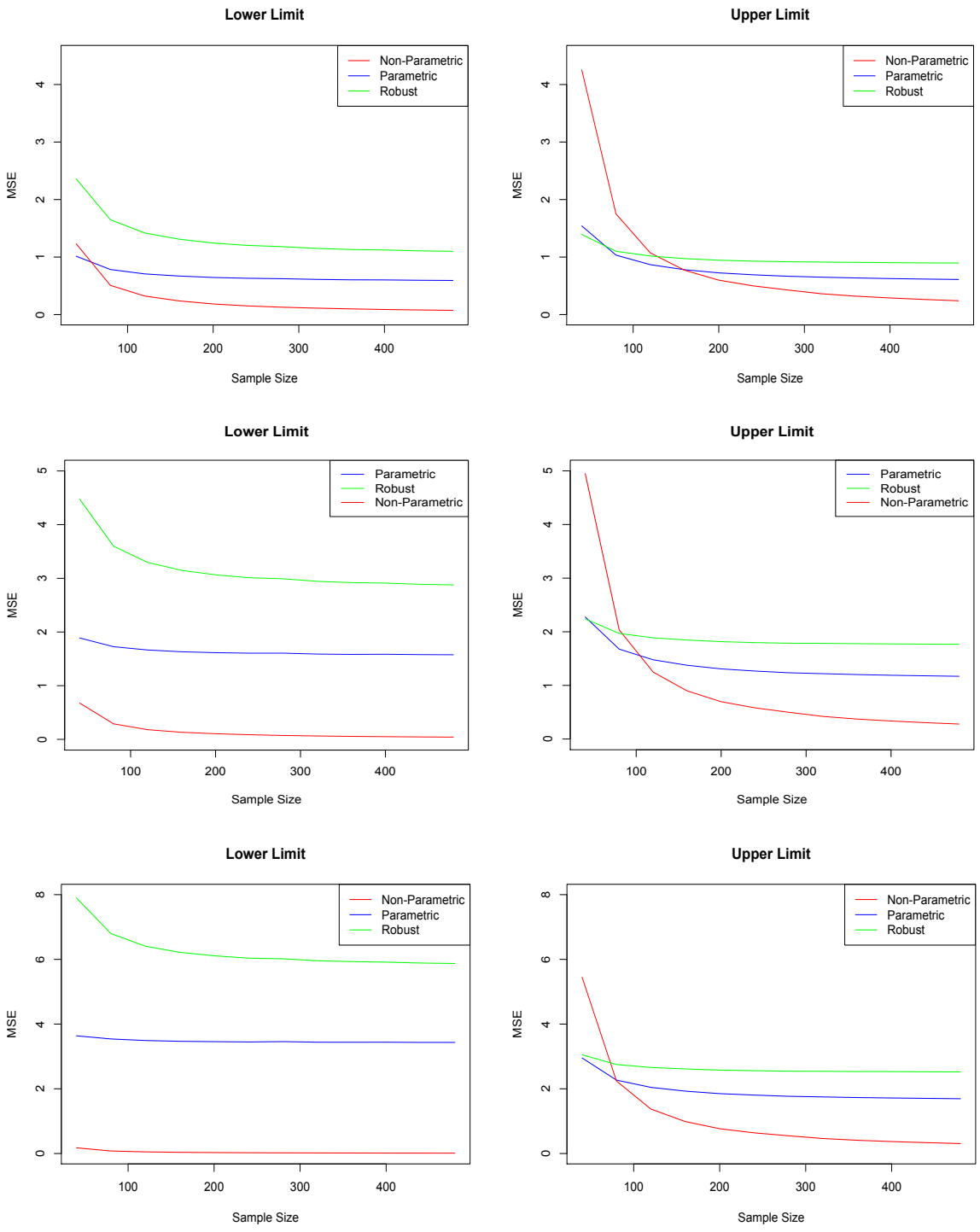

Figure 4.11: *Average empirical MSE for the three methods, where data is generated from data with skewness levels κ = 0.50 (first row), 0.75 (second row) and 0.95 (third row). The mean and variances are 20 and 9, respectively.*

Table 16: *Average empirical MSE for the three different methods, where data is generated from skew normal distributions with moderate to large levels of skewness. The mean and variance are 20 and 9, respectively.*

n	Skew = 0.50						Skew = 0.75						Skew = 0.95					
	Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	1.0150	1.5411	1.2317	4.2554	2.3598	1.3933	1.8889	2.2764	0.6768	4.9494	4.4766	2.2399	3.6381	2.9540	0.1769	5.4493	7.8973	3.0520
80	0.7835	1.0338	0.5077	1.7500	1.6487	1.1009	1.7260	1.6763	0.2888	2.0354	3.5971	1.9689	3.5408	2.2693	0.0784	2.2410	6.7984	2.7512
120	0.7072	0.8680	0.3224	1.0731	1.4191	1.0184	1.6653	1.4774	0.1808	1.2481	3.2970	1.8870	3.4929	2.0440	0.0512	1.3741	6.4077	2.6597
160	0.6703	0.7801	0.2382	0.7714	1.3113	0.9734	1.6318	1.3750	0.1320	0.8972	3.1486	1.8455	3.4681	1.9285	0.0371	0.9879	6.2187	2.6142
200	0.6451	0.7250	0.1846	0.5984	1.2434	0.9453	1.6150	1.3079	0.1050	0.6959	3.0646	1.8154	3.4572	1.8507	0.0295	0.7662	6.1109	2.5780
240	0.6312	0.6918	0.1506	0.4994	1.2028	0.9293	1.6057	1.2675	0.0866	0.5809	3.0112	1.7981	3.4474	1.8065	0.0243	0.6395	6.0377	2.5607
280	0.6239	0.6667	0.1282	0.4282	1.1819	0.9188	1.6060	1.2354	0.0724	0.4981	2.9911	1.7857	3.4584	1.7684	0.0207	0.5484	6.0189	2.5445
320	0.6116	0.6503	0.1120	0.3621	1.1510	0.9130	1.5886	1.2186	0.0630	0.4211	2.9435	1.7837	3.4401	1.7493	0.0180	0.4637	5.9577	2.5418
360	0.6042	0.6372	0.0990	0.3201	1.1320	0.9076	1.5831	1.2022	0.0552	0.3723	2.9204	1.7777	3.4394	1.7296	0.0158	0.4099	5.9317	2.5330
400	0.6026	0.6258	0.0886	0.2890	1.1234	0.9030	1.5855	1.1890	0.0502	0.3361	2.9108	1.7741	3.4403	1.7162	0.0143	0.3700	5.9147	2.5319
440	0.5966	0.6175	0.0801	0.2627	1.1084	0.8987	1.5787	1.1789	0.0454	0.3056	2.8891	1.7689	3.4337	1.7046	0.0129	0.3364	5.8871	2.5255
480	0.5933	0.6098	0.0737	0.2403	1.0994	0.8961	1.5759	1.1698	0.0416	0.2794	2.8771	1.7668	3.4334	1.6937	0.0118	0.3077	5.8733	2.5220

Overall, MSE of the estimates decreases monotonically with sample size for all three methods (Tables 15 and 16, Figures 4.9 to 4.11). Note that when $n = 40$, the robust method consistently performs the worst across all values of skewness, when estimating the lower limit (Table 16). Similarly, the non-parametric method consistently performs the worst across all values of skewness, when estimating the upper limit. This indicates that these two methods are not good choices, contrary to recommendations by the CLSI guideline. When $\kappa = 0.95$, the non-parametric method performs the best when $n \geq 80$. However, the difference in the MSE between the lower and upper limit is very noticeable. When $n = 80$, the MSE of the lower and upper limits are approximately 0.08 and 2.24, respectively. As sample size increases to 480, the MSE of the lower and upper limits decreases to approximately 0.01 and 0.31, respectively. This may be because data is concentrated to the area of the lower limit, and thus the variability of the estimates may be small. However, this is interestingly the opposite case for the robust method and warrants further investigation as to why this may be the case.

Similar to the Gaussian distributions, standardized bias (bias/σ) and standardized MSE (MSE/σ^2) were observed to be constant across all the sample sizes when skewness is held constant. This indicates that bias and MSE are directly proportional to the variability of the data, regardless of mean. Consequently, without loss of generality, results from the skew normal distribution with $\mu = 0, \sigma^2 = 1$ can be extended to skew normal distributions with any mean and variance.

Finally, consider the coverage probabilities and widths of the confidence intervals that accompany reference limits estimated by the parametric, non-parametric, and robust methods. When skewness increases slightly ($\kappa = 0.1$), the coverage probabilities decline significantly for upper limits estimated by the parametric method, and for both lower and upper limits estimated by the robust method (Table 17). Confidence intervals that accompany the lower and upper limits estimated by the non-parametric method have a coverage probability that is larger than the nominal coverage probability (0.90) on average. There is no clear relationship between sample size and coverage probability for this small level of skewness. However, as skewness increases to $\kappa = 0.25$, an inversely proportional relationship between sample size and the coverage probabilities of the parametric and robust methods appears to be established (Table 17). Coverage probabilities for lower and upper limits estimated by the non-parametric method continue to be larger than the nominal coverage probability.

Table 17*: Coverage probabilities for the three different methods, where data is generated from skew normal distributions with small levels of skewness. The mean and variance are 20 and 9, respectively.*

n	Skew = 0.1						Skew = 0.25					
	Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	0.91	0.87	N/A	N/A	0.87	0.87	0.92	0.80	N/A	N/A	0.81	0.84
80	0.91	0.86	N/A	N/A	0.85	0.89	0.88	0.79	N/A	N/A	0.74	0.80
120	0.89	0.89	0.92	0.92	0.84	0.89	0.86	0.77	0.92	0.92	0.71	0.80
160	0.89	0.86	0.93	0.93	0.83	0.87	0.81	0.74	0.94	0.92	0.64	0.73
200	0.89	0.84	0.93	0.91	0.83	0.84	0.78	0.68	0.94	0.90	0.61	0.67
240	0.90	0.86	0.96	0.95	0.84	0.86	0.76	0.69	0.93	0.95	0.57	0.67
280	0.89	0.85	0.95	0.96	0.83	0.84	0.75	0.64	0.96	0.96	0.54	0.59
320	0.87	0.84	0.94	0.92	0.81	0.84	0.67	0.61	0.92	0.92	0.48	0.57
360	0.88	0.82	0.93	0.95	0.83	0.83	0.69	0.60	0.95	0.95	0.48	0.53
400	0.87	0.80	0.93	0.92	0.81	0.80	0.65	0.53	0.93	0.92	0.43	0.48
440	0.86	0.84	0.92	0.95	0.79	0.82	0.62	0.56	0.93	0.95	0.40	0.47
480	0.87	0.80	0.92	0.90	0.80	0.76	0.60	0.50	0.93	0.90	0.37	0.43
avg	0.88	0.84	0.93	0.93	0.83	0.84	0.75	0.66	0.93	0.93	0.56	0.63

As skewness increases to higher levels ($\kappa = 0.50, 0.75, 0.95$), the coverage probabilities of confidence intervals for the parametric and robust methods significantly decline below the nominal coverage probability (Table 18). The inversely proportional relationship between coverage probability and sample size for the parametric and robust methods continues for moderate to large levels of skewness.

Table 18: *Coverage probabilities for the three different methods, where data is generated from skew normal distributions with moderate to large levels of skewness. The mean and variance are 20 and 9, respectively.*

n	Skew = 0.50						Skew = 0.75						Skew = 0.95					
	Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	0.85	0.67	N/A	N/A	0.60	0.73	0.58	0.57	N/A	N/A	0.27	0.64	0.18	0.50	N/A	N/A	0.04	0.57
80	0.68	0.60	N/A	N/A	0.39	0.63	0.24	0.45	N/A	N/A	0.06	0.46	0.01	0.34	N/A	N/A	0.00	0.37
120	0.53	0.53	0.92	0.92	0.25	0.53	0.07	0.34	0.92	0.92	0.02	0.34	0.00	0.24	0.92	0.92	0.00	0.23
160	0.44	0.46	0.93	0.92	0.18	0.42	0.03	0.27	0.93	0.92	0.00	0.23	0.00	0.17	0.93	0.92	0.00	0.15
200	0.33	0.36	0.92	0.90	0.12	0.32	0.01	0.17	0.94	0.90	0.00	0.14	0.00	0.11	0.92	0.90	0.00	0.08
240	0.22	0.34	0.94	0.95	0.06	0.27	0.00	0.14	0.95	0.95	0.00	0.09	0.00	0.08	0.94	0.95	0.00	0.05
280	0.19	0.28	0.94	0.96	0.04	0.20	0.00	0.11	0.95	0.96	0.00	0.07	0.00	0.05	0.94	0.96	0.00	0.02
320	0.13	0.28	0.93	0.92	0.03	0.19	0.00	0.09	0.93	0.92	0.00	0.04	0.00	0.03	0.93	0.92	0.00	0.01
360	0.09	0.20	0.94	0.95	0.01	0.12	0.00	0.06	0.94	0.95	0.00	0.03	0.00	0.02	0.94	0.95	0.00	0.01
400	0.07	0.17	0.90	0.92	0.01	0.10	0.00	0.03	0.93	0.92	0.00	0.01	0.00	0.01	0.93	0.92	0.00	0.00
440	0.06	0.14	0.93	0.95	0.01	0.07	0.00	0.03	0.92	0.95	0.00	0.01	0.00	0.01	0.92	0.95	0.00	0.00
480	0.04	0.11	0.91	0.91	0.01	0.06	0.00	0.02	0.90	0.91	0.00	0.01	0.00	0.01	0.93	0.91	0.00	0.00
avg	0.30	0.34	0.93	0.93	0.14	0.30	0.08	0.19	0.93	0.93	0.03	0.17	0.02	0.13	0.93	0.93	0.00	0.13

Although the coverage probabilities of parametric confidence intervals decrease as skewness increases, the widths of the confidence intervals remains approximately the same for each sample size (Tables 19 and 20). On the other hand, the widths of confidence intervals for the lower limit estimated by the non-parametric and robust methods decreases as sample size increases, and the widths of confidence intervals for the upper limits increases. This relationship corresponds with relationship between the average bias of non-parametric RIs and sample size, but is the reverse of that for robust RIs.

***Table 19**: Widths of confidence intervals for the three different methods, where data is generated from skew normal distributions with small levels of skewness. The mean and variance are 20 and 9, respectively.*

n	Skew = 0.1						Skew = 0.25					
	Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	2.66	2.66	N/A	N/A	2.78	2.83	2.66	2.66	N/A	N/A	2.76	2.90
80	1.89	1.89	N/A	N/A	1.94	1.99	1.89	1.89	N/A	N/A	1.92	2.04
120	1.54	1.54	2.75	3.15	1.58	1.61	1.54	1.54	2.53	3.45	1.59	1.66
160	1.34	1.34	2.80	3.19	1.36	1.39	1.34	1.34	2.56	3.52	1.36	1.44
200	1.19	1.19	2.12	2.41	1.21	1.24	1.19	1.19	1.97	2.64	1.22	1.28
240	1.09	1.09	2.19	2.48	1.10	1.13	1.09	1.09	1.99	2.72	1.10	1.17
280	1.01	1.01	1.88	2.15	1.02	1.04	1.01	1.01	1.73	2.35	1.02	1.08
320	0.95	0.95	1.59	1.79	0.96	0.98	0.95	0.95	1.47	1.96	0.96	1.02
360	0.89	0.89	1.66	1.86	0.90	0.92	0.89	0.89	1.52	2.04	0.90	0.95
400	0.84	0.84	1.48	1.63	0.85	0.87	0.84	0.84	1.32	1.78	0.85	0.91
440	0.80	0.80	1.37	1.58	0.82	0.83	0.80	0.80	1.28	1.73	0.82	0.86
480	0.77	0.77	1.24	1.41	0.78	0.80	0.77	0.77	1.14	1.54	0.78	0.83

Table 20: *Widths of confidence intervals for the three different methods, where data is generated from skew normal distributions with moderate to large levels of skewness. The mean and variance are 20 and 9, respectively.*

n	Skew = 0.50						Skew = 0.75						Skew = 0.95					
	Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust		Parametric		Non-Parametric		Robust	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit	Lower Limit	Upper Limit
40	2.66	2.66	N/A	N/A	2.77	3.05	2.67	2.67	N/A	N/A	2.84	3.26	2.66	2.66	N/A	N/A	2.94	3.44
80	1.89	1.89	N/A	N/A	1.96	2.18	1.89	1.89	N/A	N/A	2.02	2.34	1.89	1.89	N/A	N/A	2.11	2.48
120	1.54	1.54	2.16	3.84	1.60	1.77	1.54	1.54	1.58	4.15	1.67	1.91	1.54	1.54	0.83	4.35	1.75	2.02
160	1.34	1.34	2.13	3.92	1.38	1.54	1.34	1.34	1.59	4.23	1.43	1.65	1.34	1.34	0.84	4.44	1.50	1.76
200	1.19	1.19	1.60	2.94	1.23	1.36	1.19	1.19	1.23	3.17	1.28	1.47	1.19	1.19	0.64	3.33	1.34	1.56
240	1.09	1.09	1.66	3.03	1.13	1.25	1.09	1.09	1.25	3.27	1.17	1.35	1.09	1.09	0.66	3.43	1.23	1.44
280	1.01	1.01	1.44	2.62	1.04	1.15	1.01	1.01	1.07	2.83	1.08	1.24	1.01	1.01	0.57	2.97	1.14	1.32
320	0.94	0.94	1.23	2.19	0.98	1.09	0.94	0.94	0.92	2.36	1.02	1.17	0.94	0.94	0.49	2.48	1.06	1.25
360	0.89	0.89	1.26	2.27	0.92	1.02	0.89	0.89	0.95	2.45	0.96	1.10	0.89	0.89	0.51	2.57	1.01	1.17
400	0.84	0.84	1.11	1.98	0.87	0.97	0.84	0.84	0.85	2.14	0.90	1.04	0.84	0.84	0.45	2.24	0.95	1.11
440	0.80	0.80	1.06	1.92	0.84	0.92	0.80	0.80	0.78	2.07	0.87	1.00	0.80	0.80	0.42	2.18	0.92	1.06
480	0.77	0.77	0.97	1.71	0.80	0.89	0.77	0.77	0.71	1.85	0.83	0.95	0.77	0.77	0.38	1.94	0.88	1.01

Chapter 5

Real Data Analysis

Paediatric data has the potential to be quite complex because of the growth development stages that occur throughout childhood and adolescence. Unlike adult data, where one homogeneous group can usually represent the structure of the data across a large age range, paediatric data often require partitions into several homogeneous groups to reflect the changes in patterns associated with growth development. The collection of data can also be tedious, especially for blood analytes, as recruitment of young, healthy children is difficult. Laboratories often have to adjust for small sample sizes when estimating paediatric reference intervals (RIs) as a result of multiple partitions and limited data collection. In this section, we will be focusing on subsets of real data for three blood analytes collected by the Canadian Laboratory Initiative for Paediatric Reference Intervals (CALIPER): calcium, creatinine and alkaline phosphatase. For details regarding the collection of these data, refer to (Colantonio et al., 2012; CALIPER, 2014). Note that the RIs for each analyte are estimated for illustrative purposes only and are not intended for clinical use.

5.1 RI Estimation for Calcium

Calcium is a blood analyte associated with the endocrine system (Daniels, 2010). It is unique in the sense that it does not necessitate multiple partitions to properly reflect the nature of the data, unlike most analytes. In fact, CALIPER only used two age partitions to represent its data: 0 - < 1 year and 1 - < 19 years (Colantonio et al., 2012). The consistency of values between these two partitions can be observed in Figure 5.1.

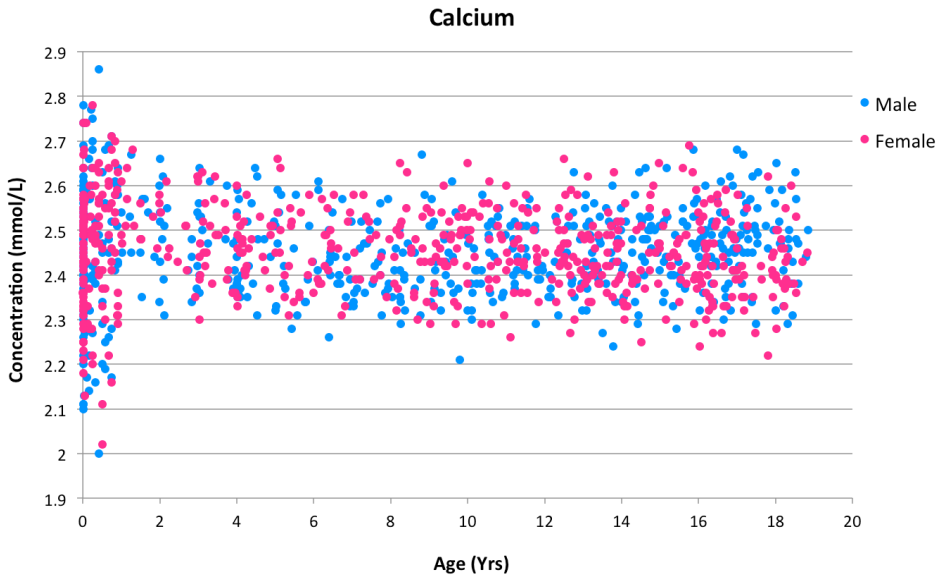

Figure 5.1: *Calcium values collected by the CALIPER group.*

Statistically speaking, the calcium data for the 1 - < 19 year age group is ideal for illustration because it appears to resemble a Gaussian distribution (Figure 5.2), but has small skewness and large sample size (Table 21). Recall that for data with small levels of skewness, the non-parametric method produced the least biased estimates for large

sample sizes (Table 13). Furthermore, although MSE for the non-parametric method was never observed to be smaller than that of the parametric and robust methods, MSE was observed to be monotonically decreasing (Figure 4.9). Thus, considering the very large sample size of the data, the non-parametric method would estimate RIs the best. The RIs (and confidence intervals) estimated by all three methods are presented in Table 22. Note that, because the sample size is large and the sample variance is very small (Table 21), the difference between the RIs estimated by all three methods is negligible.

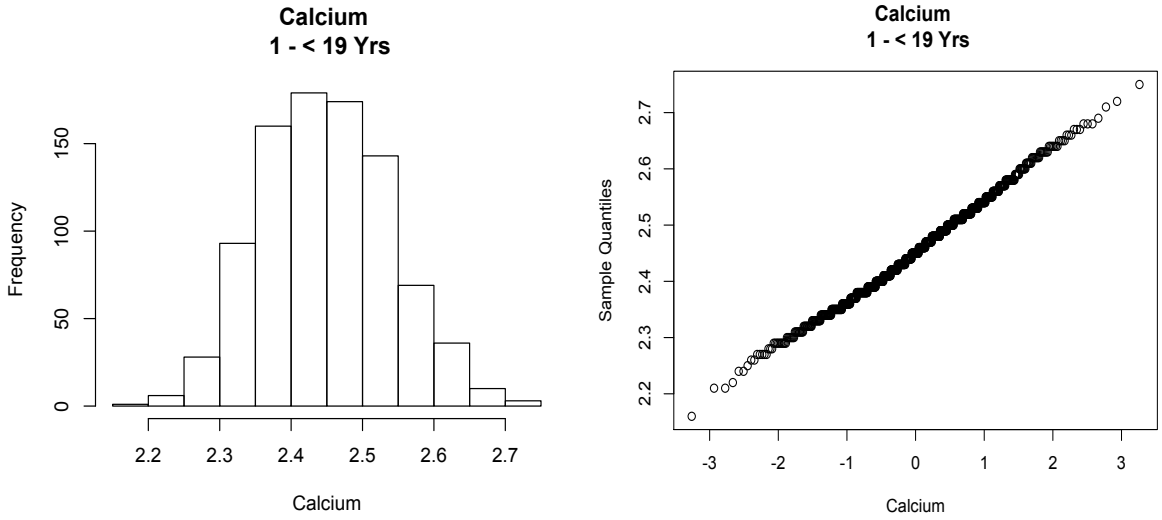

Figure 5.2: *Histogram and QQ plot of calcium values (1 - < 19 years) collected by the CALIPER group.*

Table 21: *Sample mean ($\bar{x}$), sample variance (s^2), sample size (n) and skewness (κ) of calcium values.*

Age	Gender	$\bar{x}$	s^2	n	κ
1 - < 19 yrs	F & M	2.45	0.01	902	0.16

Table 22: *RIs and confidence intervals of calcium values.*

Age	Gender	Parametric		Non-Parametric		Robust	
		Lower Limit (CI)	Upper Limit (CI)	Lower Limit (CI)	Upper Limit (CI)	Lower Limit (CI)	Upper Limit (CI)
1 - < 19 yrs	F & M	2.28 (2.27,2.29)	2.63 (2.62,2.64)	2.29 (2.28,2.30)	2.64 (2.63,2.65)	2.27 (2.27,2.28)	2.63 (2.62,2.64)

5.2 RI Estimation for Creatinine

Creatinine is a blood and urine analyte associated with the renal system (Daniels, 2010). Unlike calcium, paediatric creatinine values fluctuate with age and gender because of its dependency on muscle mass (Daniels, 2010). Creatinine data collected by CALIPER and tested through the enzymatic method is displayed in Figure 5.3.

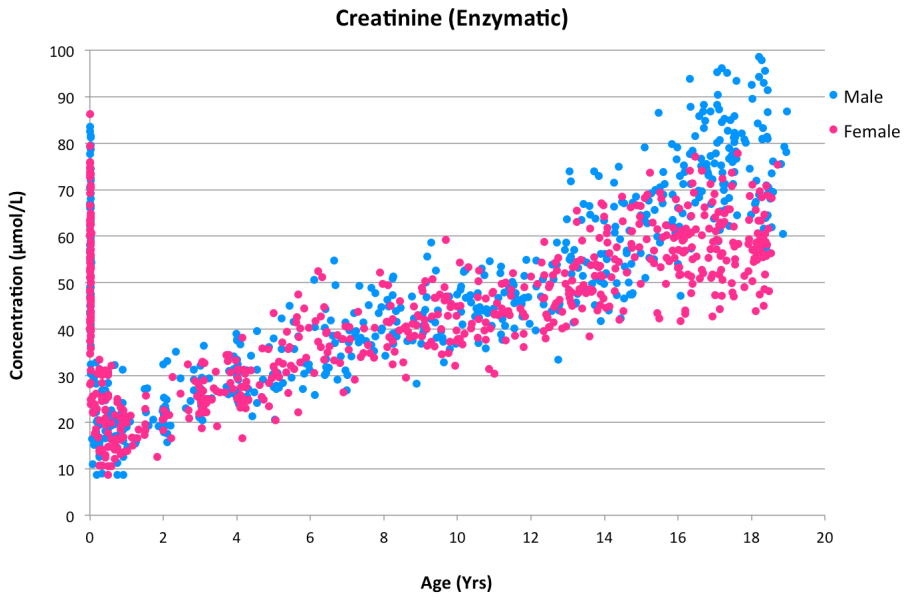

Figure 5.3: *Creatinine values collected by the CALIPER group.*

The general increasing trend presented by the creatinine values makes it difficult to divide the data into homogeneous groups (Figure 5.3). CALIPER partitioned this data into 6 age groups and for one of these age groups, gender was separated in order to properly represent the trends in the data: 0 - 14 days, 15 days - < 2 years, 2 - < 5 years, 5 - < 12 years, 12 - < 15 years, 15 - < 19 years (females), and 15 - < 19 years (males) (Colantonio et al., 2012). Some of these partitions appear to closely resemble a Gaussian distribution, such as the 15 - < 19 year old males, as shown in Figure 5.4. Other partitions, such as the 15 - < 19 year old females, resemble skewed distributions, as shown in Figure 5.5.

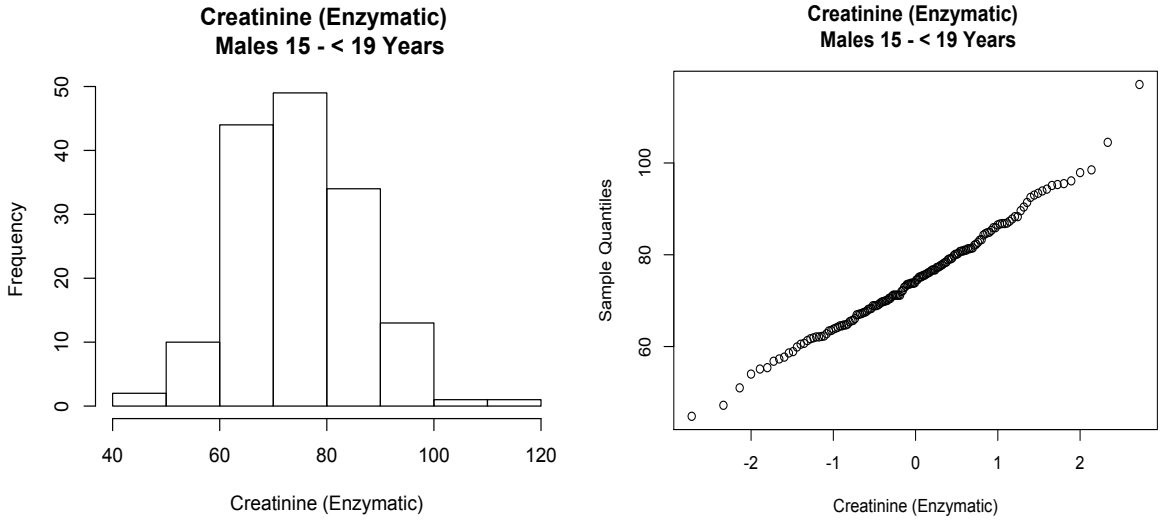

Figure 5.4: *Histogram and QQ plot of male creatinine values (15 - < 19 years) collected by the CALIPER group.*

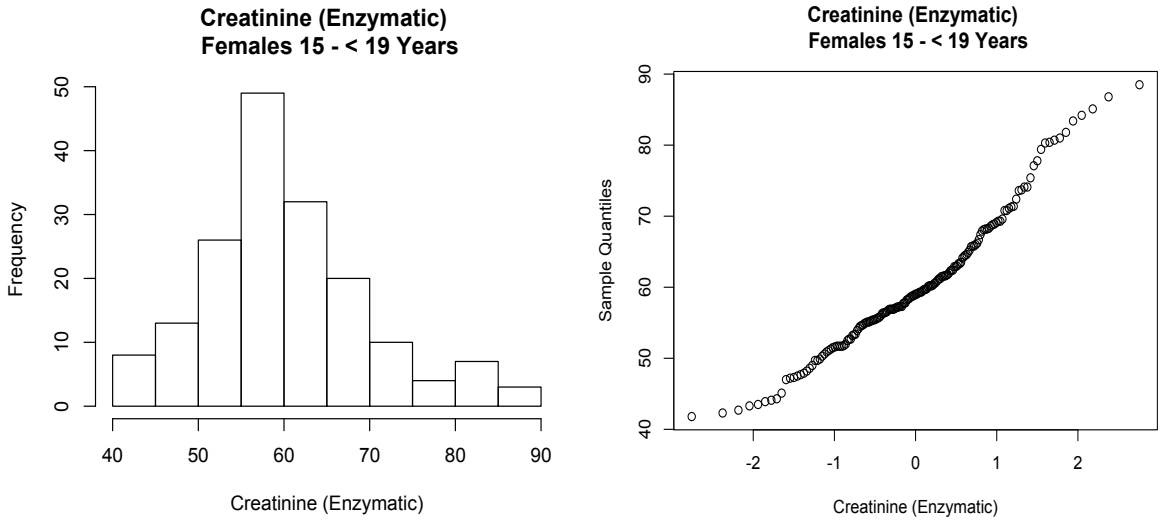

Figure 5.5: *Histogram and QQ plot of female creatinine values (15 - < 19 years) collected by the CALIPER group.*

The sample mean ($\bar{x}$), sample variance (s^2), sample size (n) and skewness (κ) of creatinine values of females and males aged 15 - < 19 yrs are provided in Table 23. Note that the creatinine values for the 15 - < 19 year old females has a moderate to large skewness. Based on the results of our simulation study, it is clear that the non-parametric would perform the best in estimating RIs for this level of skewness and moderate sample size (Tables 14 and 16). Also, although the creatinine values for the 15 - < 19 year old males displayed features of a Gaussian distribution in its histogram and QQ plot (Figure 5.4), the data has small to moderate positive skewness. Note that the close resemblance to a Gaussian distribution leads discrepancies between normality tests. For example, the Kolmogorov-Smirnov test rejects normality for this data, but the Shapiro-Wilk test does not. Thus, if a laboratory was dealing with data with similar characteristics, the choice of

the method may not be clear. However, if the choice was based on sample size and skewness, our simulation study showed that the non-parametric method would produce the least biased estimates (Tables 13 and 14). However, the consistency of the estimates produced by the non-parametric method is not as good as the parametric method, especially for the upper limit of a RI.

Table 23: *Sample mean ($\bar{x}$), sample variance (s^2), sample size (n) and skewness (κ) of creatinine values.*

Age	Gender	$\bar{x}$	s^2	n	κ
15 - < 19 yrs	F	60.21	91.62	172	0.66
15 - < 19 yrs	M	74.82	132.70	154	0.35

The RIs (and confidence intervals) estimated by all three methods are presented in Table 24. The difference between the upper limits of the RIs for the male data estimated by the non-parametric and parametric methods is minimal and up for clinical interpretation. In this case, the larger MSE for the non-parametric method only seems to be reflected by the much wider confidence intervals for the limits, compared to the parametric method.

Table 24: *RIs and confidence intervals of creatinine values.*

Age	Gender	Parametric		Non-Parametric		Robust	
		Lower Limit (CI)	Upper Limit (CI)	Lower Limit (CI)	Upper Limit (CI)	Lower Limit (CI)	Upper Limit (CI)
15 - < 19 yrs	F	41.5 (39.4,43.5)	79.0 (76.9,81.0)	43.3 (41.8,45.1)	84.2 (80.4,88.5)	39.8 (37.7,41.7)	78.0 (75.7, 80.6)
15 - < 19 yrs	M	52.2 (49.6,54.9)	97.4 (94.8,100.0)	54.0 (44.8,57.3)	97.9 (95.1,117.1)	51.43 (48.4, 54.4)	97.16 (94.1, 100.0)

5.3 RI Estimation for Alkaline Phosphatase

Alkaline phosphatase is a blood analyte associated with the hematological system (Daniels, 2010). Like creatinine, paediatric alkaline phosphatase values fluctuate with age and gender. For the alkaline phosphatase values collected by CALIPER (Figure 5.6), data was partitioned into 7 age groups, and for 3 of the age groups, data was further partitioned into gender groups: 0 - 14 days, 15 days - < 1 year, 1 - < 10 years, 10 - < 13 years, 13 - < 15 years (females), 13 - < 15 years (males), 15 - < 17 years (females), 15 - < 17 years (males), 17 - < 19 years (females), and 17 - < 19 years (males).

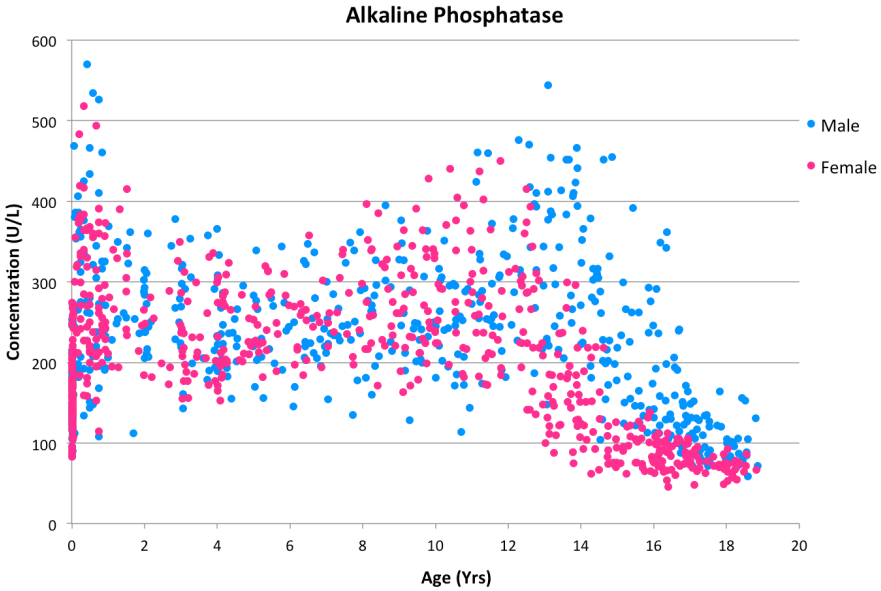

Figure 5.6: *Alkaline phosphatase values collected by the CALIPER group.*

Due to the large number of partitions, groups within the 13 - < 19 year range have sample sizes of less than 120. Now, the Clinical Laboratory Standards Institute (CLSI) guideline would recommend that the robust method should be used to estimate RIs for these small datasets. However, would the parametric method be better in estimating the RIs for these datasets since data should be transformed to a symmetric distribution when applying the robust method? An interesting dataset to look at is the alkaline phosphatase values from females 17 to less than 19 years of age. It appears to be slightly skewed (Figure 5.7), yet normality tests do not reject normality of this dataset.

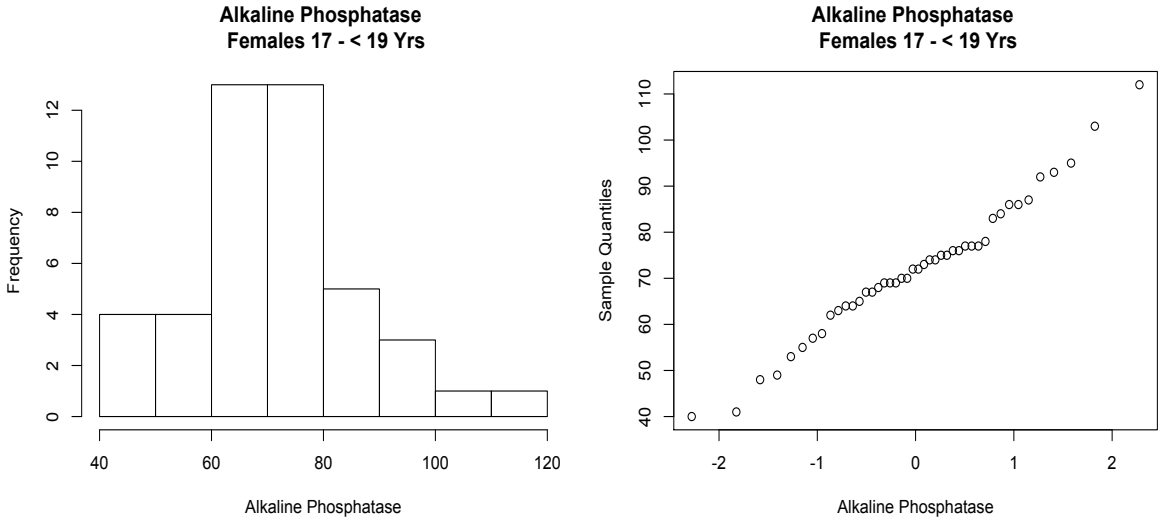

Figure 5.7: *Histogram and QQ plot of female alkaline phosphatase values (17 - < 19 years) collected by the CALIPER group.*

The sample mean ($\bar{x}$), sample variance (s^2), sample size (n) and skewness (κ) of alkaline phosphatase values of females aged 17 - < 19 years are provided in Table 25. The alkaline values for females aged 17 - < 19 years has small positive skewness. In addition, the sample size is very small. For this level of skewness and sample size, the parametric method estimated the lower limit with the least amount of bias while the robust method estimated the upper limit with the least amount of bias (Table 13). The difference between the two methods in terms of MSE was small (Table 15).

Table 25: *Sample mean ($\bar{x}$), sample variance (s^2), sample size (n) and skewness (κ) of alkaline phosphtase values.*

Age	Gender	$\bar{x}$	s^2	n	κ
17 - < 19 yrs	F	71.93	223.32	44	0.2

The RIs (and confidence intervals) estimated by all three methods are presented in Table 26. The difference between the actual lower and upper limits is quite small, but it is up to laboratories to interpret the significance of this difference, noting that the upper limit estimated by the parametric method leads to false positives on average, and the upper limit estimated by the robust method leads to false negatives on average. If the difference is clinically significant for this analyte, then laboratories may wish to have the final RI (42.6, 102.0), where the lower limit is estimated by the parametric method, and the upper limit is estimated by the robust method.

Table 26: *RIs and confidence intervals of alkaline phosphatase values.*

Age	Gender	Parametric		Non-Parametric		Robust	
		Lower Limit (CI)	Upper Limit (CI)	Lower Limit (CI)	Upper Limit (CI)	Lower Limit (CI)	Upper Limit (CI)
15 - < 19 yrs	F	42.6 (36.2,49.0)	101.2 (94.8,107.6)	40 (N/A)	112 (N/A)	41.1 (33.9,48.3)	102.0 (95.0,109.1)

Chapter 6

Discussion

When data comes from a Gaussian distribution, it is clear that the parametric method is the best approach to estimate reference intervals (RIs) in the sense that it produces estimates with the least bias and mean squared error (MSE) across all sample sizes compared to the non-parametric and robust methods. Nevertheless, when dealing with data that has large variance and small sample size, the parametric method slightly overestimates the lower limit and underestimates the upper limit, possibly leading to false positives for abnormality. Note that the occurrence of these false positives is in addition to the 5% error in misclassification of healthy individuals when 95% RIs are used. However, the occurrence of additional false positives is clinically negligible since the bias of the estimates produced by the parametric method is within the variability of the data. Therefore, it is not a concern for clinicians.

In addition, not only does the parametric method produce asymptotically unbiased and consistent estimates of RIs, but the confidence intervals that accompany its estimates are also statistically sound. The coverage probability of the confidence intervals is close to the nominal level, and the width of the confidence intervals is the smallest out of the three methods. Furthermore, the simplicity of this method, combined with its applicability

to small sample sizes, makes this method attractive as a whole. The lack of attention that this method gets from the Clinical Laboratory Standards Institute (CLSI) guideline is startling, given all the advantages it has to offer, when data comes from a Gaussian distribution and also when data is slightly skewed. Fortunately, the systematic review presented in Chapter 3 and (Daly et al., 2013) has found that the parametric method is still being used in practice.

Unfortunately, choosing the best method for data with considerable, moderate and large skewness is not as clear as it is for data from Gaussian distributions. Firstly, one must note that although data may be skewed, normality tests such as the Anderson-Darling, Pearson, and Shapiro tests, may not reject normality because of low power associated with such tests. In these cases, laboratories may follow recommendation to use the parametric method because they are under the impression that they are dealing with Gaussian data. We ran a simulation to examine the level and power of these tests with data generated from Gaussian and skew normal distributions, where skewness (κ) = 0.10, 0.25, 0.50, 0.75, and 0.95. When data was generated from a Gaussian distribution, these tests performed well. However, when data was generated from a skew normal distribution, where $\kappa = 0.10$, these tests produced a lot of false negatives (>88%). Nevertheless, if a laboratory proceeded to estimate RIs when $\kappa = 0.10$, the parametric method still provides the best estimates when sample size is small, as shown by our simulation results (Chapter 4). Therefore, lack of power is not a concern in this case. When data was generated from a skew normal distribution, where $\kappa = 0.95$, these tests

demonstrated high power with large sample size. However, for small sample size (n = 40), the tests failed to reject normality at an alarmingly high rate (Anderson-Darling: 37%, Pearson: 59%, and Shapiro: 32%). For this level of skewness and sample size, the simulation results in Chapter 4 show that the parametric method did not produce good estimates. Thus, laboratories should also look at QQ plots in addition to conducting normality tests when evaluating the normality of their data.

Another complication of skewed data is, when we observed different levels of skewness, no method estimated RIs with the least amount of bias across all samples sizes. Moreover, it was often the case that one method estimated the lower limit with the least amount of bias, while another method estimated the upper limit with the least amount of bias. To our knowledge, it has never been suggested to use one method to estimate the lower limit of a RI and another method to estimate the upper limit. However, the importance of producing the least biased RIs may sway laboratories to estimate RIs in this manner. This is especially true for a particular laboratory test where a false negative result for a particular laboratory test may lead to a patient missing life saving treatment, and/or a false positive result may lead to a patient undergoing unnecessary invasive and costly treatment. Alternatively, if it is more favourable to increase false negatives in order to decrease false positives (or vice versa) due to the cost and/or nature of further screening and treatment (or lack thereof) that these results can lead to for a particular laboratory test, then one of the three methods may serve this purpose for both limits.

Along with guidelines outlining the best methods in various circumstances, laboratories need to be advised that when they are estimating RIs with highly variable data, it is important to use a larger sample size in order to obtain more precise estimates.

Note that the skewed distributions considered in the simulation are positive. Only positively skewed distributions were examined because most distributions of laboratory test values are positively skewed, perhaps because the results cannot fall below a value of 0 or a minimum limit of detection. However, if laboratories do deal with negatively skewed distributions, we would expect the above results of the simulation for the lower limit to be that of the upper limit and vice versa. Thus, observations noted in the results of positively skewed distributions, such as the non-parametric method estimating the lower limit better then the upper limit, would be reversed for negatively skewed distributions.

In addition, transformations of skewed distributions to symmetric distributions were not considered in the simulation study, as the observations gathered from the results of Gaussian scenarios (and scenarios with slightly skewed data) most likely apply in these cases. As mentioned in Chapter 2, it has been suggested that data should be transformed to a symmetric distribution before applying the robust method to estimate RIs. This might explain the poor performance of the method with skewed distributions. However, with Gaussian distributions (and hence symmetric), the parametric method outperformed the robust method and will be recommended in these cases going forward.

In the future, it would be of interest to investigate the possibility of a bias adjusted parametric approach. Since the bias of parametric method has been derived in this thesis,

this should not be difficult to implement, and the performance of the method could be examined through a simulation study similar to this one. In addition, this simulation only examined the performances of the methods given that the data has already been partitioned. Partitioning is another critical step in establishing RIs, and is particularly crucial with paediatric data. There is a less subjective, automatic approach to assist laboratories in determining partitions and should be developed. Alternatively, partitioning could be avoided altogether with RI estimation methods involving reference curves. Thus, the performances of methods available to produce reference curves should be investigated in the future.

Bibliography

Adibi A, Haghighi M, Hashemipour M, et al. (2012). New reference values for thyroid volume by ultrasound in Semirom, Iran: report of a pilot study. *Pakistan Journal of Medical Sciences.* **28**(2),321-323.

Azzalini A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*. **32**(2),159-188.

CALIPER. (2014). http://www.caliperdatabase.com [Online; accessed June-2014].

Cinaz P, Yesilkaya E, Onganlar YH, et al. (2012). Penile anthropometry of normal prepubertal boys in Turkey. *Acta Paediatrica*. **101**(1),e33-36.

Clifford SM, Bunker AM, Jacobsen JR, et al. Age and gender specific pediatric reference intervals for aldolase, amylase, ceruloplasmin, creatine kinase, pancreatic amylase, prealbumin, and uric acid. *Clinica Chimica Acta*. **412**(9-10),788-790.

Clinical and Laboratory Standards Institute (CLSI). (2008). *Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline*. CLSI document C28-A3. Clinical and Laboratory Standards Institute, Wayne. Third Edition.

Colantonio DA, Kyriakopoulou L, Chan MK, et al. (2012). Closing the gaps in pediatric laboratory reference intervals: a CALIPER database of 40 biochemical markers in a healthy and multiethnic population of children. *Clinical Chemistry*. **58**(6),854-868.

Cole TJ, Green PJ. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine.* **11**(10),1305-1319.

Daly CH, Liu X, Grey VL, et al. (2013). A systematic review of statistical methods used in constructing pediatric reference intervals. *Clinical Biochemistry*. **46**(13-14),1220-1227.

Daniels R. (2010). *Delmar's guide to laboratory and diagnostic tests*. Delmar, Cengage Learning, Clifton Park. Second Edition.

Dixon WJ. (1953). Processing data for outliers. *Biometrics* **9**(1),74-89.

Dunn OJ. (1961). Multiple comparisons among means. *Journal of the American Statistical Association.* **56**(293),52-64.

Efron B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics, Philadelphia.

Efron B, Tibshirani R. (1993). *An introduction to the bootstrap.* Chapman & Hall Inc., New York.

Goh SY, Aragon JM, Lee YS, et al. (2011). Normative data for quantitative calcaneal ultrasound in Asian children. *Annals Academy of Medicine Singapore.* **40**(2),74-79.

Graham RL, Knuth DE, Patashnik O. (1994). *Concrete mathematics: a foundation for computer science*. Addison-Wesley Publishing Company, Reading.

Green A, Morgan I, Gray K. (2003). *Neonatology and laboratory medicine*. ACB Venture Publications, London. Second Edition.

Harris EK, Boyd JC. (1990). On dividing reference data into subgroups to produce separate reference ranges. *Clinical Chemistry*. **36**(2),265-270.

Higgins JPT, Green S, eds. (2011). *Cochrane handbook for systematic reviews of interventions*. 5.1.0 Edition. http://www.cochrane-handbook.org [Online; accessed May-2012].

Horn PS, Pesce AJ. (2005). *Reference intervals: a user's guide*. AACC Press, Washington.

Hulecki LR, Small SA. (2011). Behavioral bone-conduction thresholds for infants with normal hearing. *Journal of the American Academy of Audiology*. **22**(2),81-92.

KiGGS. (2014). http://www.kiggs-studie.de/english/home.html [Online, accessed June-2014].

Lahti A, Hyltoft Petersen P, Boyd JC, et al. (2002). Objective criteria for partitioning Gaussian-distributed reference values into subgroups. *Clinical Chemistry*. **48**(2),338-352.

LOOK Lifestyle Study. (2014). http://www.look.org.au/ [Online; accessed June-2014].

Mann HB, Whitney DR. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*. **18**(1):50-60.

Pavlov IY, Wilson AR, Delgado JC. (2012). Reference interval computation: which method (not) to choose? *Clinica Chimica Acta*. **413**(13-14),1107-1114.

Pediatric Reference Intervals. (2014). http://www.aruplab.com/pediatrics/about/pediatric-reference-intervals [Online; accessed June-2014].

Reed AH, Henry RJ, Mason WB. (1971). Influence of statistical method used on the resulting estimate of normal range. *Clinical Chemistry*. **17**(4),275-284.

Royston P, Wright EM. (1998). A method for estimating age-specific reference intervals ('normal ranges') based on fractional polynomials and exponential transformation. *Journal of the Royal Statistical Society: Series A*. **161**(1),79-101.

Sen A, Srivastava M. (1990). *Regression analysis: theory, methods, and applications*. Springer-Verlag New York Inc., New York.

Sidak Z. (1967). Rectangular confidence regions for the means of multivariate distributions. *Journal of the American Statistical Association.* **62**(318),626-633.

Sinton TJ, Cowley DM, Bryant SJ. (1986). Reference intervals for calcium, phosphate, and alkaline phosphatase as derived on the basis of multichannel-analyzer profiles. *Clinical Chemistry*. **32**(1),76-79.

Solberg HE. (2006). Establishment and use of reference values. In: Burtis CA, Ashwood ER, Bruns DE, eds. *Tietz textbook of clinical chemistry and molecular diagnostics*. St. Louis: Elsevier Saunders, St. Louis. Fourth Edition.

Tamimi W, Tamim H, Felimban N, etal. (2011). Age- and gender-specific reference intervals for serum lipid levels (measured with an Advia 1650 analyzer) in school children. *Pediatrics International.* **53**(6),814-819.

Tamimi W, Tamim H, Felimban N, et al. (2012). Age- and gender-specific reference intervals for serum glucose levels in school children measured by an Advia 1650 chemistry analyzer. *Journal of Pediatric Biochemistry*. **2**(2),101-108.

Tukey JW. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company, Reading.

Uemura O, Honda M, Matsuyama T, et al. (2011). Age, gender, and body length effects on reference serum creatinine levels determined by an enzymatic method in Japanese children: a multicenter study. *Clinical and Experimental Nephrology*. **15**(5),694-699.