THE PERFORMANCE AND ROBUSTNESS OF CONFIDENCE INTERVALS FOR THE MEDIAN OF A SYMMETRIC DISTRIBUTION CONSTRUCTED ASSUMING SAMPLING FROM A CAUCHY DISTRIBUTION

by

JENNIFER YUE CAO

M.S., Kansas State University, USA, 2007

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Approved by:

Major Professor
Paul Nelson

Abstract

Trimmed means are robust estimators of location for distributions having heavy tails. Theory and simulation indicate that little efficiency is lost under normality when using appropriately trimmed means and that their use with data from distributions with heavy tails can result in improved performance. This report uses the principle of equivariance applied to trimmed means sampled from a Cauchy distribution to form a discrepancy function of the data and parameters whose distribution is free of the unknown median and scale parameter. Quantiles of this discrepancy function are estimated via asymptotic normality and simulation and used to construct confidence intervals for the median of a Cauchy distribution. A nonparametric approach based on the distribution of order statistics is also used to construct confidence intervals. The performance of these intervals in terms of coverage rate and average length is investigated via simulation when the data are actually sampled from a Cauchy distribution and when sampling is from normal and logistic distributions. The intervals based on simulation estimation of the quantiles of the discrepancy function are shown to perform well across a range of sample sizes and trimming proportions when the data are actually sampled from a Cauchy distribution and to be relatively robust when sampling is from the normal and logistic distributions.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would love to thank my research advisor Dr. Paul Nelson and would love to show my great appreciation for his close guidance. Without his effort and help, this work would have been very difficult. I want to express my gratitude to my supervisory committee, Dr.Weixing Song and Dr. Kun Chen for their advice and support. I also want to show my acknowledgement to all the faculty, the department staff and graduate students at the department of Statistics. I thank the department of Statistics and Kansas State University for the opportunity given to me to pursue study here.

I would love to thank all the members in my family in China. Their unselfish love and support is the main drive for me to have survived the difficulties I have met in my life and study.

In addition, I would like to extend my appreciation to all the friends I have made in Manhattan. Thank them for their help with my study and life in here. Their friendship and understanding will be memorized in my heart forever.

Chapter 1 Introduction

1.1 Goal of the Report

Although the heavy tails of a Cauchy distribution can make it difficult to work with, there are several parametric methods based on trimmed means for constructing confidence intervals for its median, denoted by μ , from data consisting of a random sample. Using simulation, this report studies, compares and assess the performance of these methods, and a nonparametric one, in terms of actual coverage rate and interval length when sampling is from Cauchy, normal and logistic distributions. A study of confidence intervals based on the maximum likelihood estimator of μ for the Cauchy distribution is left to future work.

1.2 Introduction to Cauchy Distribution

The Cauchy–Lorentz distribution, named after Augustin Cauchy and Hendrik Lorentz, is a symmetric, heavy tailed, continuous probability distribution. Among statisticians it is known as the Cauchy distribution, while among physicists, it is known as the Lorentz distribution, Lorentz(ian) function, or Breit–Wigner distribution. The Cauchy distribution has the probability density function of the form

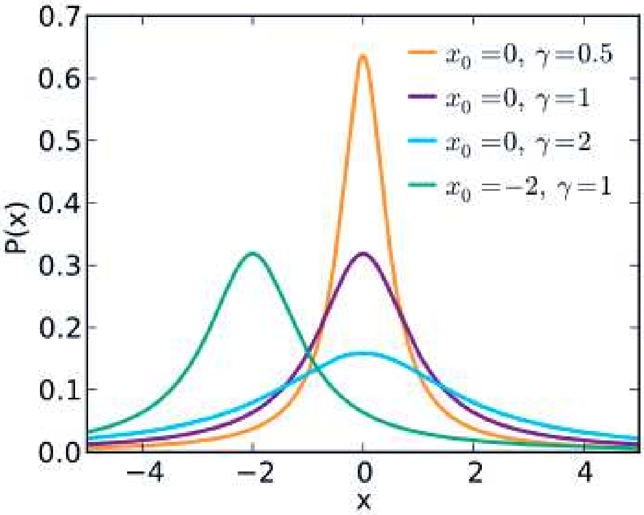
$$f(x) = (1/\sigma)g((x-\mu)/\sigma)), \quad (1)$$

where

$$g(z) = 1/\pi(1+z^2) \qquad\qquad (2)$$

is the standard Cauchy density, and μ and $\sigma > 0$ are respectively location and scale parameters, taken here to be unknown. Note that μ is the median of (1). Figure 1.1 illustrates some representative Cauchy densities. The one in purple is the standard Cauchy density. Cauchy densities look similar to normal densities. However, they have much heavier tails, so heavy that they do not have a mean. The mean and standard deviation of the Cauchy distribution are undefined. The practical consequence of this is that collecting 1,000 data points gives no more accurate an estimate of the mean than does a single point. The high rate at which the Cauchy density produces outliers makes it useful for studying the robustness of statistical procedures with respect to extremes in particular and departures from normality in general.

Figure1.1 Graphs for Cauchy Densities with $x_0 = \mu \quad \gamma = \sigma$



The standard Cauchy distribution is a *t*-distribution with one degree of freedom and hence the distribution of the ratio of two independent, standard normal random variables.

In physics, economics, mechanics, and electrical engineering, and many other technical and scientific fields, especially in dealing with calibration problems, Cauchy distributions are used instead of normal distributions when extreme values are comparatively likely to occur. Heavy tails are also described by Jacob and Protter(1998) as leading to what is called *fat tailed behavior*. The Cauchy distribution is often used for counter-examples in probability theory. Specifically, as noted above, it does not possess the usual descriptive moments such as mean and variance. As noted above, it does have a median, denoted μ here, which is also the mode. This extreme behavior motivates my report, a study of the performance of confidence intervals for μ based on trimmed means.

Let $\mathbf{X}_n = \{X_i, i = 1, 2, ..., n\}$ be iid having the Cauchy density given in (1). Having observed $\mathbf{X}_n = \mathbf{x}_n$, suppose that it is desired to construct a *nominal* $1-\alpha$ confidence intervals for the median μ of the form

$$CI(\alpha, \mathbf{x}_n) = [L(\mathbf{x}_n), U(\mathbf{x}_n)]$$

where

$$P(\mu \in CI(\alpha, X_n)) = CR \approx 1-\alpha$$

and *CR* denotes the actual coverage rate.

The Cauchy distribution, because of its heavy tail, often results in data sets that contain both large and small, extreme outliers,which adversely affect the performance of the sample mean $\overline{X}_n = \sum_{i=1}^{n} X_i / n$ as an estimator of the median μ. Since $\overline{X}_n$ has the same distribution as X_1, it also has median μ . However, it also follows that: (i) if we sample n = 10,000 observations, $\overline{X}_n$ is no better as an estimator of μ than is a single observation X_1; (ii) commonly used confidence intervals based on the approximate normality of $\overline{X}_n$ are not even applicable. Although the sample median is inefficient, according to Fisher and Tilanus(1964), it is the simplest consistent estimator of μ and is used in practice. The maximum likelihood estimator of μ is consistent and asymptotically efficient. But it is difficult to compute and interpret, as pointed out in Fisher and Tilanus(1964). This report uses another consistent family of estimators of μ, trimmed means.

1.3 Introduction to Trimmed Means

A trimmed mean is computed by discarding a certain percentage of the lowest and the highest values in a sample and then computing the mean of the remaining observations. For many years, the trimmed mean has been an extremely popular estimator of location parameters, as noted in Stigler(1973). Theory and simulation indicate that little power is lost under normality when using appropriately trimmed means can result in substantially higher power than tests of hypothesis based on the sample mean when sampling from a heavy-tailed distribution. Using the trimmed mean, Yang(2001) achieved robustness of parameter estimation in a zero-inflated Poisson model, in which excessive zeroes occur. Walfish(2006) gave another example of applications of trimmed means to accommodate recent changes in the Olympic scoring system for ice skating, In this study, trimmed means, of which the sample median is a special case, are used to estimate the median μ of a Cauchy distribution to remedy the deficiencies of $\bar{X}_n$ described above. Specifically, suppose that it is desired to trim the 100p% largest and smallest observation, $0 \le p < 1/2$, and average the rest. Specifically, letting $\{X_{(i)}; i = 1,2,...,n\}$ denote the order statistics obtained from $\{X_i; i = 1,2,...,n\}$, a trimmed mean for specified integer *r* is defined by

$$\bar{X}_{p,n} = \sum_{i=r+1}^{n-r} X_{(i)} / (n-2r),$$

where, approximately in some cases, p = r/n. Note that for *r* = 0, $\bar{X}_n = \bar{X}_{0,n}$. The trimmed mean $\bar{X}_{p,n}$ is the sample median if: (i) *n* = 2*m* + 1 and *r* = *m*; (ii) *n* = 2*m* and *r* = *m*-1. We assume that the amount of trimming is fixed prior to analyzing the data and that sampling is from a distribution symmetric about its median. Trimmed means are examples of equivariant estimators. Equivariance, as described below, facilitates the construction of confidence intervals for the location parameter μ from location-scale families such as the Cauchy given in (1).

Chapter 2 Confidence Intervals Based on Equivariance in Location-Scale Models

2.1 Equivarinace

I begin with the definition of equivariance and then apply it to use trimmed means to construct confidence intervals for the median of a Cauchy distribution. The setting here is the Cauchy family given in (1), but it could be used for any location-scale family.

<u>Definition</u>: Equivariance: Let $\hat{\mu}(\underline{\mathbf{X}})$ be an estimator of μ and $\hat{\sigma}(\underline{\mathbf{X}})$ an estimator of σ for the family of densities given in (1) such that for any constants $a > 0$ and b

$$\hat{\mu}(a\underline{\mathbf{X}}+b\underline{\mathbf{1}}) = \mathrm{a}\,\hat{\mu}(\underline{\mathbf{X}}) + \mathrm{b}, \qquad (3)$$

$$\hat{\sigma}(a\underline{\mathbf{X}}+b\underline{\mathbf{1}}) = \mathrm{a}\,\hat{\sigma}(\underline{\mathbf{X}}).$$

Estimators satisfying the first line in (3) include $\hat{\mu}(\underline{\mathbf{X}})$ equal to a trimmed mean and those satisfying the second line include

$$\hat{\sigma}_1(X_n) = \sum_{i=1}^{n} \left| X_i - \hat{\mu}(X_n) \right| / n \quad , \qquad (4)$$

o

$$\hat{\sigma}_2(X_n) = \sqrt{\sum_{i=1}^{n} (X_i - \hat{\mu}(X_n))^2 / n} \quad , \qquad (5)$$

or

$$\hat{\sigma}_3(X_n) = (X_{3/4} - X_{1/4})/2 \quad , \qquad (6)$$

where $X_{3/4}$ is the sample third quartile, $X_{1/4}$ is the sample first quartile.

To use equivariance, note that for X having a Cauchy distribution given in (1), $Z = (X-\mu)/\sigma$ has the standard Cauchy density given in (2). Hence, letting

$\mathbf{Z}_n = (Z_1, Z_2, ..., Z_n),$ using equivariant estimators of location and scale, we have that

$$T_n(\mathbf{X}) \equiv (\hat{\mu}(\underline{X}) - \mu)/\hat{\sigma}(\underline{X})$$

$$= \hat{\mu}(\underline{Z}_n)/\hat{\sigma}(\underline{Z}_n)$$

$$= T_n(\mathbf{Z}_n)$$

has a distribution function, denoted $H_n(\bullet)$, free of the unknown μ and σ. Specifically $H_n(z)$ = $P(T_n(Z) \le z)$ can be computed from (2) without knowing μ or σ.

2.2 Constructing Confidence Intervals Using Equivariance

To use the setup given above to construct a confidence interval for μ**,** find quantiles $t_{\alpha/2,n}, t_{1-\alpha/2,n}$ so that $H(t_{\alpha/2,n}) = \alpha/2$ and $H(t_{1-\alpha/2,n}) = 1-\alpha/2$. Note again that these quantiles do not depend on the unknown location and scale parameters. Then,

$$P(t_{\alpha/2,n} \le T_n \le t_{1-\alpha/2,n}) = 1-\alpha,$$

and hence, having observed $\underline{\mathbf{X}}_n$ = $\underline{\mathbf{x}}_n$, an exact $1-\alpha$ confidence interval for μ is given by

$$[\hat{\mu}(\underline{x}) + t_{\alpha/2,n}\hat{\sigma}(\underline{x}), \hat{\mu}(\underline{x}) + t_{1-\alpha/2,n}\hat{\sigma}(\underline{x})] \qquad (6)$$

In this report, I will use the trimmed mean $\overline{X}_{p,n} = \sum_{i=r+1}^{n-r} X_{(i)}/(n-2r)$, an equivariant estimator of the median of the Cauchy distribution, and $\hat{\sigma} = \hat{\sigma}_3(X_n) = (X_{3/4} - X_{1/4})/2$. Then, an exact $1-\alpha$ confidence interval for μ is given by

$$[\overline{x}_{p,n} + t_{\alpha/2,n}\,\hat{\sigma}_3(\underline{x}), \overline{x}_{p,n} + t_{1-\alpha/2,n}\,\hat{\sigma}_3(\underline{x})] \ .$$

Instead of computing the quartiles $t_{\alpha/2,n}, t_{1-\alpha/2,n}$, which would be very difficult, two methods will be used in this report to estimate them.

2.3 Methods for Estimating the Quartiles of $H(\bullet)$

Method I: Asymptotic Normality

Note that equivariance allows us, without loss of generality, to take $\mu = 0$ and $\sigma = 1$.

Letting $n \to \infty$ and $r \to \infty$ so that $r/n \to p$, Serfling(2001) showed that $\hat{\sigma}_3(X_n)$ is a consistent estimator of σ. With $k = 1/2 - p$, $p \neq 1/2$, Rothenberg et. al. (1964) showed that, in distribution, as $n \to \infty$,

$$W_n = \sqrt{(kn)}(\bar{X}_{p,n} - \mu)/\sigma \to N(0, \gamma^2) \qquad (7)$$

where, $\gamma^2 = [(1-k)\tan^2(\pi k/2)/k + 2\tan(\pi k/2)/\pi k - 1]$, for $n = 2m + 1$, $m = 0, 1, \ldots, ;$ $r = m - [nk] - 1$. Since $[nk]/n = (m-1)/n - p$ so that approximately $k = ½ - p$.

Then, in distribution, as $n \to \infty$, Slutsky's Theorem, as in Hogg(2004), yields

$$\sqrt{kn}T_n = (\sigma / \overset{\wedge}{\sigma}_3(X_n))W_n$$
$$\to N(0, \gamma^2).$$

Hence, for large *n*, approximately

$$t_{\alpha/2,n} = -\gamma z_{\alpha/2} / \sqrt{(nk)} \quad \text{and} \quad t_{1-\alpha/2,n} = \gamma z_{\alpha/2} / \sqrt{nk} \ ,$$

where $\Phi(z_\delta) = \delta$, $0 < \delta < 1/2$ and $\Phi(\cdot)$ denotes the standard normal distribution function.

Henceforth, I will fix *p* and let *k* = ½-*p*.

The method described above needs to be modified to handle the case when $p = 1/2$ so that the trimmed mean is the sample median. Specifically, the trimmed mean $\overline{X}_{p,n}$ is the sample median if: (i) $n = 2m + 1$ and $r = m$; (ii) $n = 2m$ and $r = m$-1. As $n \to \infty$, Ferguson et.al (1996) showed that for sample medians $\{m_n\}$ of the Cauchy distribution,

$$\sqrt{n}(m_n - \mu) \to N(0, \pi^2\sigma^2/4)$$

Again, using Slutsky's Theorem, a large sample, approximate $1-\alpha$ confidence interval for μ is given by

$$[m_n - z_{\alpha/2,n}\hat{\sigma}(\underline{\mathbf{x}})\pi/(2\sqrt{n}), m_n + z_{\alpha/2,n}\hat{\sigma}(\underline{\mathbf{x}})\pi/(2\sqrt{n})] \qquad (8)$$

Method II: *Simulation*

For large *R*, generate $\{\mathbf{Z}_i^* = (Z_{ij}^*, j = 1,2,...,n), i = 1,2,...,R\}$, where $\{Z_{ij}^*\}$ are iid random variables having the standard Cauchy distribution in (2) and compute $\{T_n(Z_i^*), i = 1,2,..., R\}$. Let $\hat{H}_n(\bullet)$ be the empirical distribution function obtained from $\{T_n(Z_i^*), i = 1,2,..., R\}$, defined by $\hat{H}_n(z) = \#\{j;\ T_n(Z_j^*) \le z\}/R$. Letting $\hat{H}_n^{-1}(\beta)$ be the corresponding estimate of the order β quantile, then approximately,

$$\hat{t}_{\alpha/2,n} = \hat{H}_n^{-1}(\alpha/2) \text{ and } \hat{t}_{1-\alpha/2,n} = \hat{H}_n^{-1}(1-\alpha/2)$$

Chapter 3 A Nonparametric Confidence Interval

For the sample median of any continuous distribution, such as the Cauchy, having a unique median, we have that *P(X<μ)=P(X>μ)=1/2*. Then, based on a random sample of size *n*, *Y*= #{*i*, *Xi*<μ}has a Binomial(*n*,1/2) distribution. Letting $\{X_{(i)}\}$ denote the order statistics and '*r*' an index such that $P(X_{(r)} < \mu < X_{(n-r+1)} = P(Z_{(r)} < 0 < Z_{(n-r+1)}) = 1-\alpha$, a $1-\alpha$ confidence interval for μ is given by:

$$[X_{(r)}, X_{(n-r+1)}].$$

To approximate the index *r* needed for constructing an approximate, nonparametric $1-\alpha$ CI, for large sample size *n* =2*m*+1, using approximate normality,

$$\begin{aligned} P(X_{(r)} \le \mu \le X_{(n-r+1)}) &= P(r \le Y \le n-r) \\ &= \sum_{k=r}^{n-r} \binom{n}{k} (\frac{1}{2})^n \cong \Phi(\frac{n-r-n/2+1/2}{\sqrt{n/4}}) - \Phi(\frac{r-n/2-1/2}{\sqrt{n/4}}) \\ &= \Phi(\frac{m-r}{\sqrt{n/4}}) - \Phi(\frac{r-m}{\sqrt{n/4}}) \end{aligned}$$

Thus, we have $\frac{m-r}{\sqrt{n/4}} \cong z_{\alpha/2}$, and thus $r \cong m - z_{\alpha/2}\sqrt{n/4}$. Since r is an integer, we take $r = [m - z_{\alpha/2}\sqrt{n/4}]$, where $[\bullet]$ is the greatest integer function. Henceforth, intervals constructed using this approach will be referred to as *Method III.*

Chapter 4 Simulation Study

4.1 Simulation Algorithm

I used nominal coverage rate $1-\alpha$ = 0.95 and selected representative values of n = $2m$+ 1 and p. For each such choice I generated data from the Cauchy, normal and logistic distributions and used the following algorithm:

(1) Generate R = 1000 independent random samples $\{z_i^* = (z_{i1}, z_{i2}, ..., z_{in})\}$, i= 1,2,…, R, each consisting of n values independently sampled from (2). As described above, use sample quantiles to estimate $t_{\alpha/2,n}, t_{1-\alpha/2,n}$.

(II) Independently generate n observations $\mathbf{x}^*$ from (2). Construct confidence intervals using the three methods. Record whether or not each interval contains $\mu = 0$ and its length.

(III) Independently generate n observations $\mathbf{x}^*$ from the standard Cauchy, scaled standard normal and logistic distributions. Construct confidence intervals using the three methods for each of the three data sets. Record whether or not each interval contains $\mu = 0$ and its length.

(IV) Independently repeat (II)-(III) N =1000 times.

Assess and compare the performance of the confidence intervals across all parameter settings.
I begin the summary of my results using Cauchy data.

4.2 Cauchy Data Method I: (Asymptotic Normality)

I carried out the algorithm given above using the asymptotic normality method for samples from the standard Cauchy distribution with parameter settings *R*=1000, *n*=11, 21, 31, 101 and *p*=0.1, 0.2, 0.3, 0.5. First, for $p < 1/2$, which *p*=*r/n, k*=1/2-*p*, I computed γ given by the formula $\gamma^2 = [(1-k)\tan^2(\pi k/2)/k + 2\tan(\pi k/2)/\pi k - 1]$ and used it to find estimates of the critical points $t_{\alpha/2,n} = -\gamma z_{\alpha/2}/\sqrt{(nk)}$ and $t_{1-\alpha/2,n} = \gamma z_{\alpha/2}/\sqrt{nk}$. Then, I independently generated a set of 1000 samples (*N*=1000) from the standard Cauchy distribution and used this value, along with trimmed means and $\hat{\sigma}_3$, to construct Method I nominal 0.95 confidence intervals for each sample. For $p = 1/2$, I used Serfling's asymptotic variance given in (8). The results are summarized in terms of estimated coverage rates, average and median confidence interval lengths from Cauchy data in Table 4.1 below. The SAS code I used is given in Appendix A.

4.3 Cauchy Data Method II: (Simulation)

I carried out the algorithm given above for samples from the standard Cauchy distribution with parameter settings *R*=1000, *n*=11, 21, 31, 101 and *p*=0.1, 0.2, 0.3, 0.5. Given sample size *n*, I set the index *r* so that the trimming proportions were approximately 0.1, 0.2, 0.3 and 0.5. Specifically, for instance, for *n*=31, *r*=3, which gives *p*=3/31=0.1, I generated 1000 samples of size *n* = 31 from the standard Cauchy distribution and deleted the three largest and smallest observations and averaged the remaining 25 observations from each sample to yield 1000 trimmed means. I sorted these 1000 trimmed means and found the 2.5th and 97.5th sample quartiles needed for a 0.95 confidence interval. Due to the symmetry of the Cauchy density,

$t_{\alpha/2,n} = -t_{1-\alpha/2,n}$. Hence I used ± the average of absolute values of the sample quartiles as the critical points. Then, I independently generated another set of 1000 samples (*N*=1000) from the standard Cauchy distribution and used this method to construct nominal 0.95 confidence intervals for such as *n* =31 and *r* = 3, which gives *p*= 3/31=0.1 for each sample. The simulation results are summarized in terms of estimated coverage rates, average and median confidence interval lengths in Table 4.2 below. The SAS code I used is given in Appendix B.

Cauchy Data Method I

Table 4.1 Simulated Coverage Rates, Average Lengths Based on Asymptotic Normality

		p=0.1	*p*=0.2	*p*=0.3	*p* =0.5
n=11	Estimated t	0.912	0.896	0.892	N/A
	Coverage Rate	62.2	86.3	93.6	95.5
	Average CI Length	2.66	2.61	2.6	2.71
	Median of CI Length	2.21	2.17	2.17	2.25
n=21	Estimated t	0.659	0.648	0.646	N/A
	Coverage Rate	66.8	84.3	89.3	91
	Average CI Length	1.36	1.33	1.33	1.38
	Median of CI Length	1.26	1.24	1.23	1.28
n=31	Estimated t	0.542	0.533	0.532	N/A
	Coverage Rate	76.1	91.5	95	95
	Average CI Length	1.23	1.2	1.2	1.25
	Median of CI Length	1.18	1.57	1.15	1.2
n=101	Estimated t	0.3	0.295	0.2946	N/A
	Coverage Rate	81.2	91.1	94.1	94.6
	Average CI Length	0.6	0.59	0.59	0.62
	Median of CI Length	0.599	0.59	0.588	0.61

Cauchy Data Method II

Table 4.2 Estimated Coverage Rates, Average Lengths Based on Equivariance Cauchy Data

		p =0.1	p =0.2	p =0.3	p =0.5
n=11	Estimated t	1.216	0.918	0.878	0.887
	Coverage Rate	94.7	95.7	95.4	94.4
	Average CI Length	3.55	2.68	2.5	2.59
	Median of CI Length	2.95	2.23	2.13	2.15
n=21	Estimated t	1.056	0.795	0.748	0.799
	Coverage Rate	93.5	95.9	95.7	95.3
	Average CI Length	2.17	1.64	1.54	1.64
	Median of CI Length	2.02	1.52	1.43	1.53
n=31	Estimated t	0.762	0.558	0.53	0.573
	Coverage Rate	95.6	95.8	95.6	95.6
	Average CI Length	1.72	1.26	1.2	1.3
	Median of CI Length	1.65	1.21	1.15	1.24
n=101	Estimated t	0.419	0.325	0.304	0.305
	Coverage Rate	93.4	94.7	94.8	94.6
	Average CI Length	0.843	0.654	0.611	0.614
	Median of CI Length	0.836	0.648	0.606	0.608

4.4 Cauchy Data Method III: (Nonparametric)

The results from the method III are summarized in terms of estimated coverage rates, average and median confidence interval lengths in Table 4.3 below. The SAS code is in Appendix C.

Cauchy Data Method III

Table 4.3 Estimated Coverage Rates, Average Lengths Based on Nonparametric Method

	n=11	*n*=21	*n*=31	*n*=101
Coverage Rate	100	99.2	98.7	98.7
Average CI Length	52.08	2.86	1.83	1.83
Median of CI Length	14.41	2.61	1.75	1.75

4.5 Assessments of Coverage Rates Method I and II Cauchy Data.

Using the data in Tables 4.1 and 4.2, estimated coverage rates (*CR*) were compared for the first two methods in terms of sample size, *n*, and trimming proportion (trimp), *p*. Proc GPLOT of SAS was used to make plots to visualize *CR* and average length in terms of *n*, *p* and Method. Proc GLM was used to fit a linear model to the data in the Tables 4.1 and 4.2 with *CR* and average interval length as the responses and *n*, method, *p* as predictors. Since the observed CR's fell in a narrow range, it was not deemed necessary to use logistic regression. The linear model is expressed as:

$$Y = \beta_0 + \beta_1 n + \beta_2 p + \beta_3 method + \beta_4 n \times p + \beta_5 n \times method + \beta_6 p \times method + \beta_7 n \times p \times method + \beta_8 p^2 + \varepsilon$$

Y: Coverage Rate or Average Length;

$\beta_1 - \beta_8$: Parameters to be estimated;

ε: Residuals term;

Model assumption: $\varepsilon \overset{IID}{\sim} N(0, \sigma^2)$

In these two models, all the variables are treated continuous except that method is treated as a categorical variable.

The output from Proc GLM using *CR* as the response from the Cauchy data summary tables, is given in Table 4.4 below, guided my plots and conclusions. The coefficient of determination R^2 = 0.84 and the residuals plot and the Q-Q plot (Figure 4.1a) and b)) indicate that the model fits the data reasonably well. All of the effects involving just method and /or trimming proportion are statistically significant at the 0.05 level. Most of the variation is accounted for by effects involving just these two sources. Somewhat surprisingly, effects involving sample size *n* account for just around 3 % of the total. The following graphical presentations of the data in Figures 4.2-4.4, highlighting main effects and two way interactions, plotted separately for different values of the third effect, help in interpreting the regression analysis.

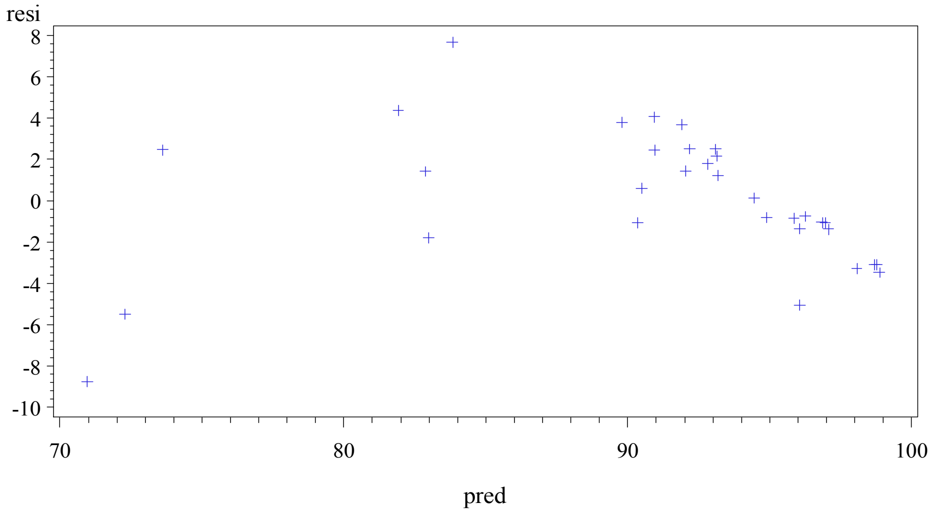
Table 4.4 Output of Proc GLM using CR as Response for Cauchy data

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1807.627116	225.953390	14.95	<.0001
Error	23	347.695071	15.117177		
Corrected Total	31	2155.322187			

R-Square	Coeff Var	Root MSE	CR Mean
0.838681	4.278053	3.888081	90.88438

Source	DF	Type III SS	Mean Square	F Value	Pr > F
N	1	54.4201484	54.4201484	3.60	0.0704
Method	1	660.7717734	660.7717734	43.71	<.0001
Trimp	1	398.8731387	398.8731387	26.39	<.0001
N*method	1	79.5145073	79.5145073	5.26	0.0313
N*trimp	1	28.2600714	28.2600714	1.87	0.1848
trimp*method	1	317.9783047	317.9783047	21.03	0.0001
N*trimp*method	1	36.5160714	36.5160714	2.42	0.1338
trimp*trimp	1	240.8263718	240.8263718	15.93	0.0006

Figure 4.1 a) Residuals Plot CR Cauchy Data



b) Normality QQ plot CR Cauchy Data

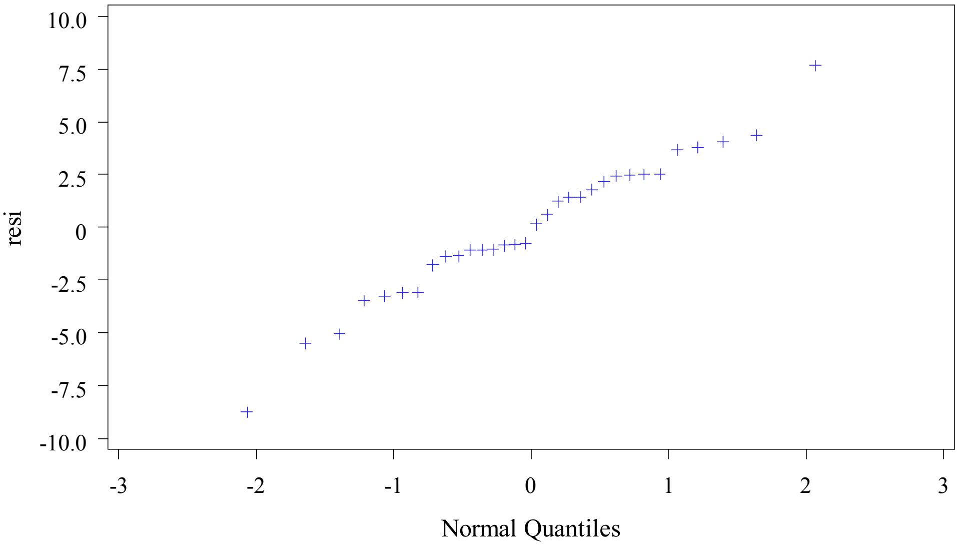
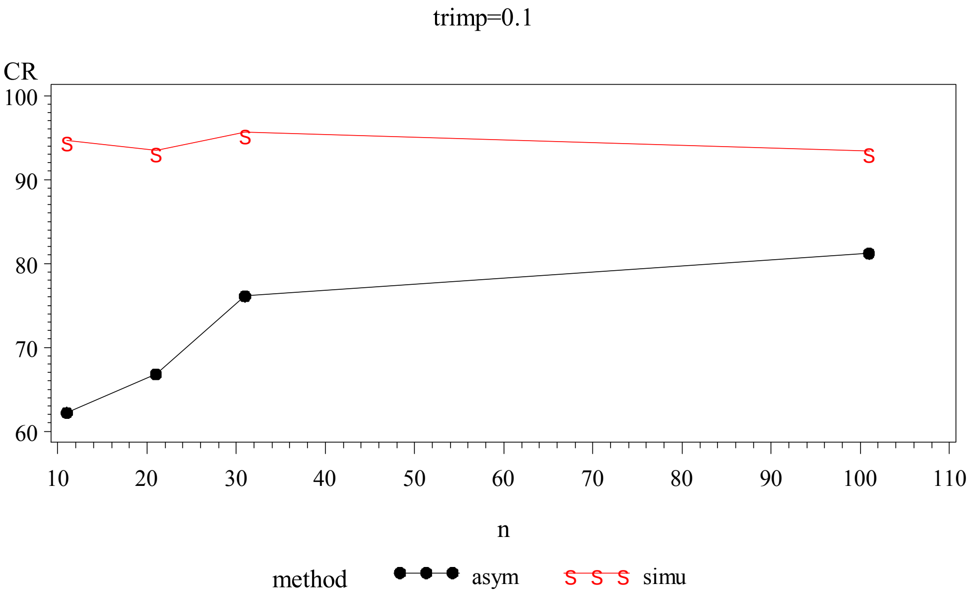
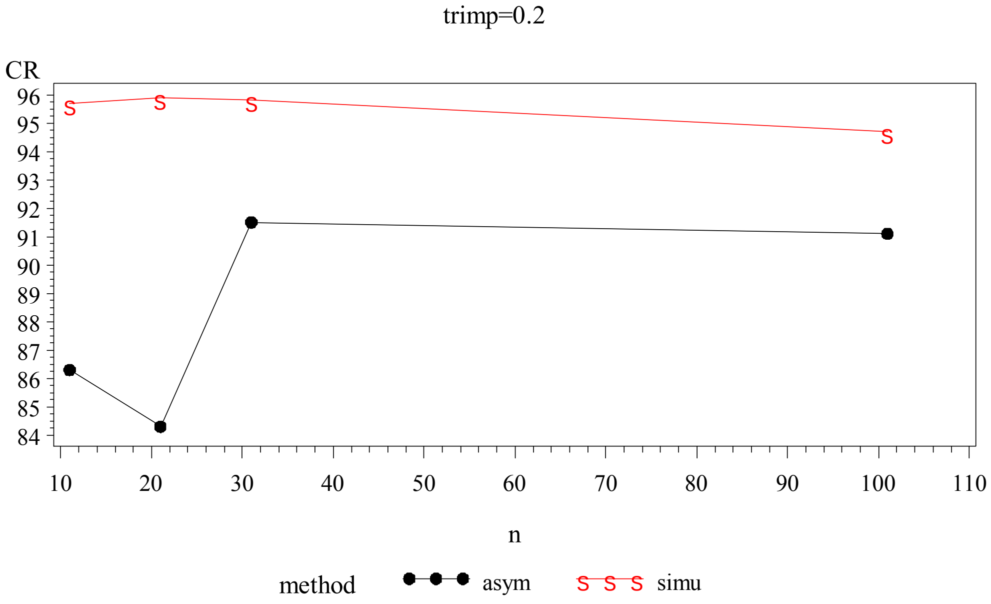


Figure 4.2 CR: Sample Size by Method Interaction Cauchy Data

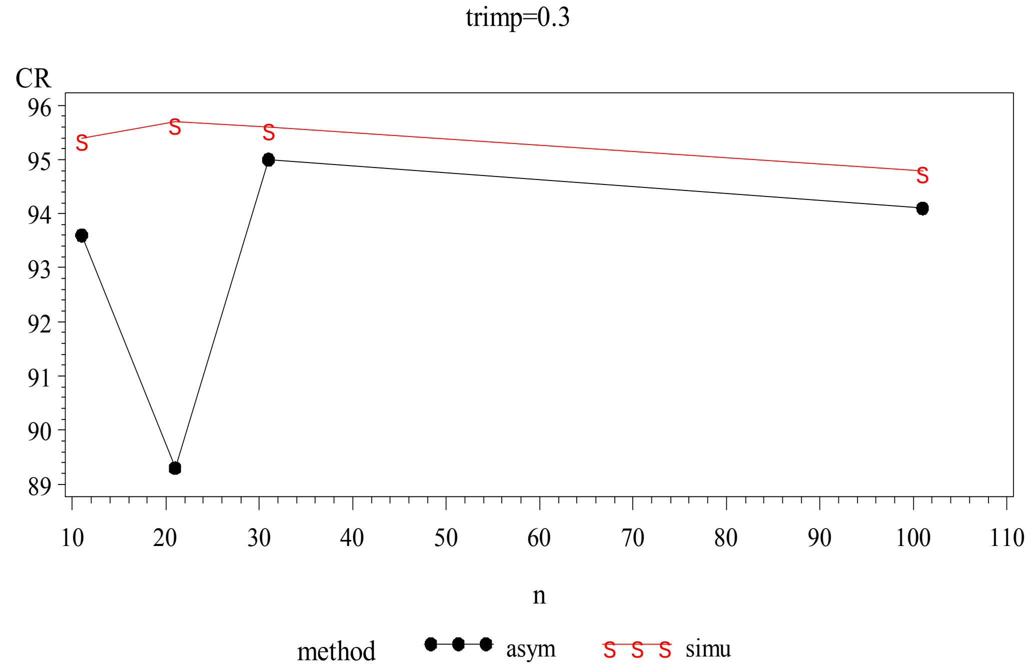
a) *p* = 0.1



b)*p*=0.2



c) *p* =0.3


trimp=0.3
CR
96
95
94
93
92
91
90
89
10
20
30
40
50
60
70
80
90
100
110
n
method
asym
simu

d) *p* =0 .5

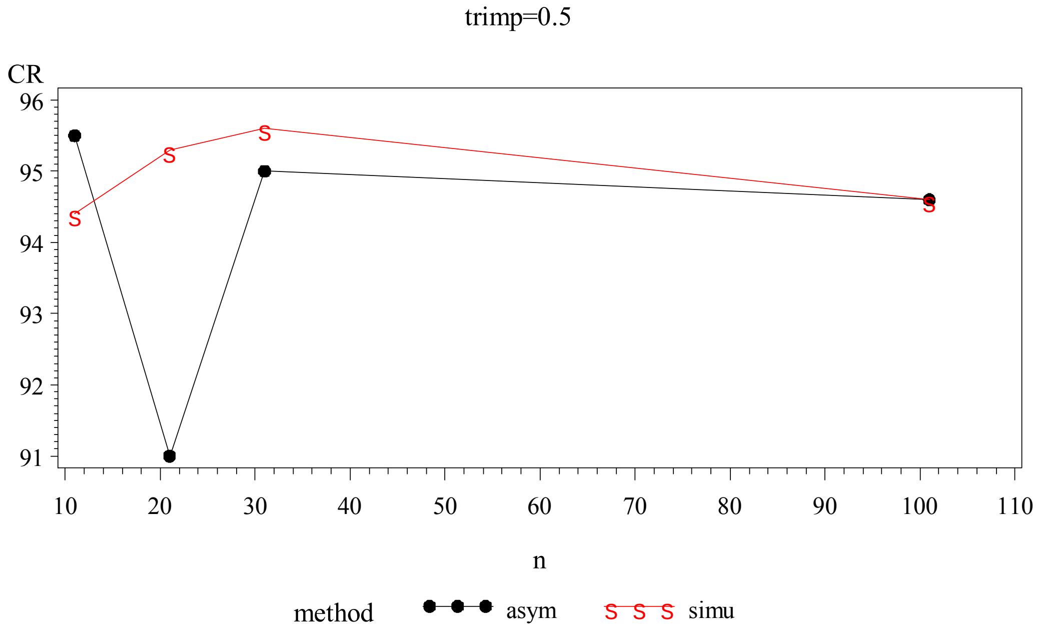
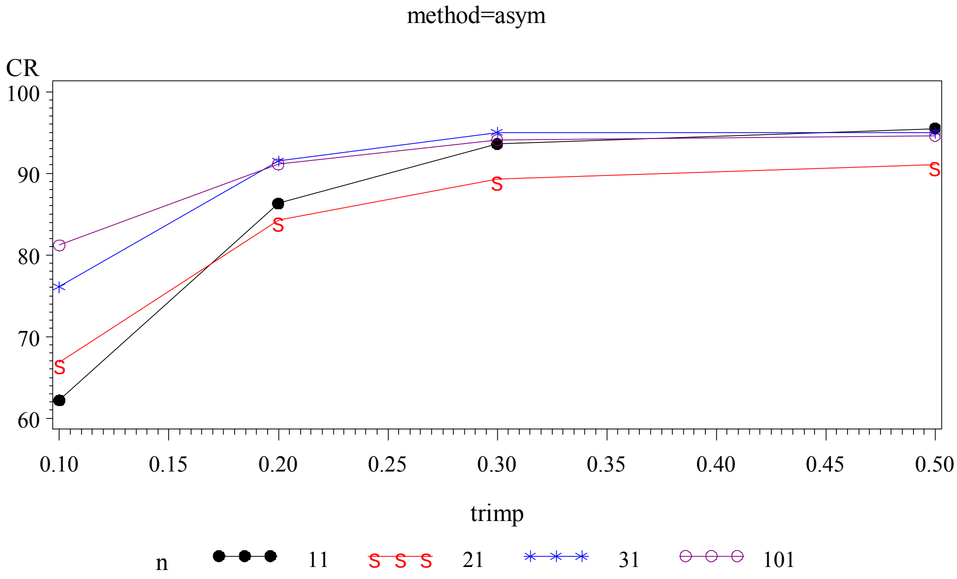

trimp=0.5
CR
96
95
94
93
92
91
10
20
30
40
50
60
70
80
90
100
110
n
method
asym
simu

Figure 4.3 CR: Trimming Proportion by Sample Size Interaction Cauchy Data

a) Method I



b) Method II

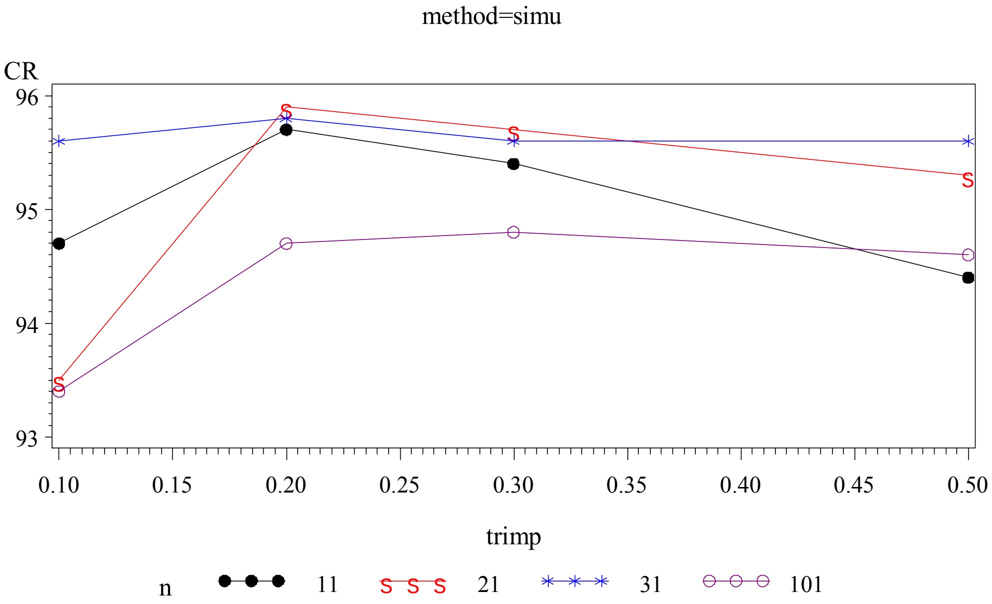
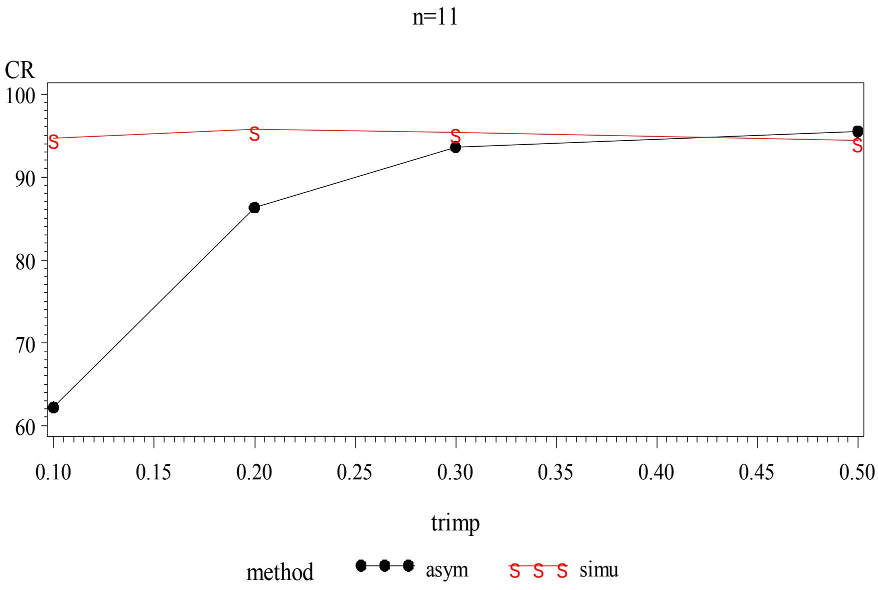
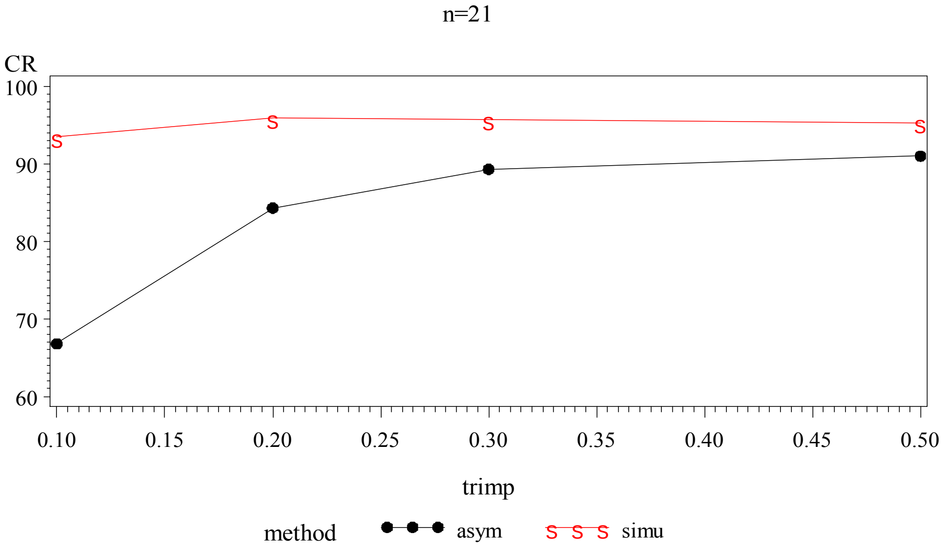


Figure 4.4 CR: Trimming Proportion by Method Interaction Cauchy Data

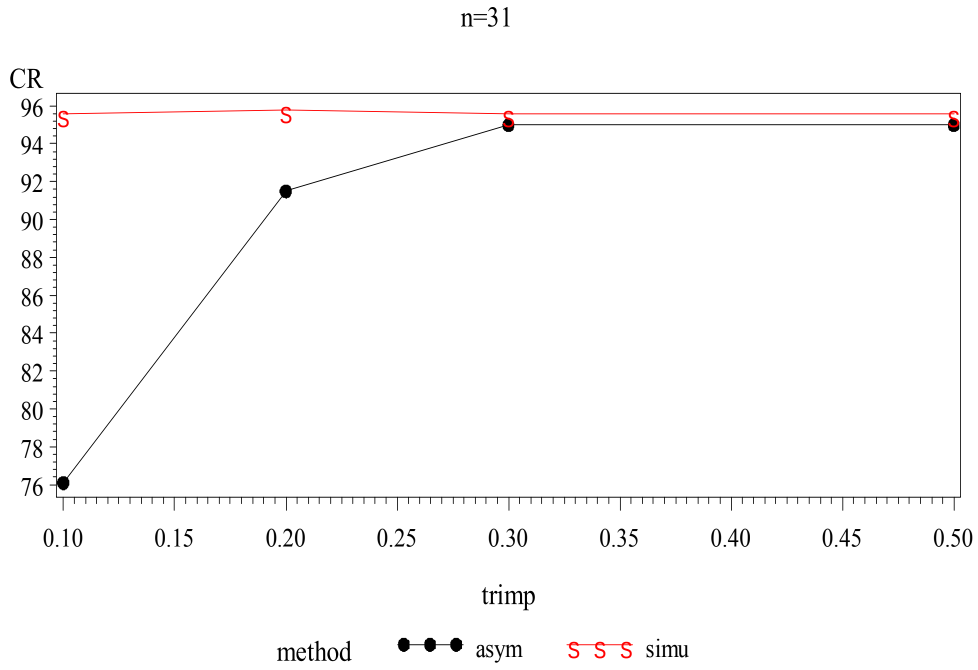
a) *n*=11



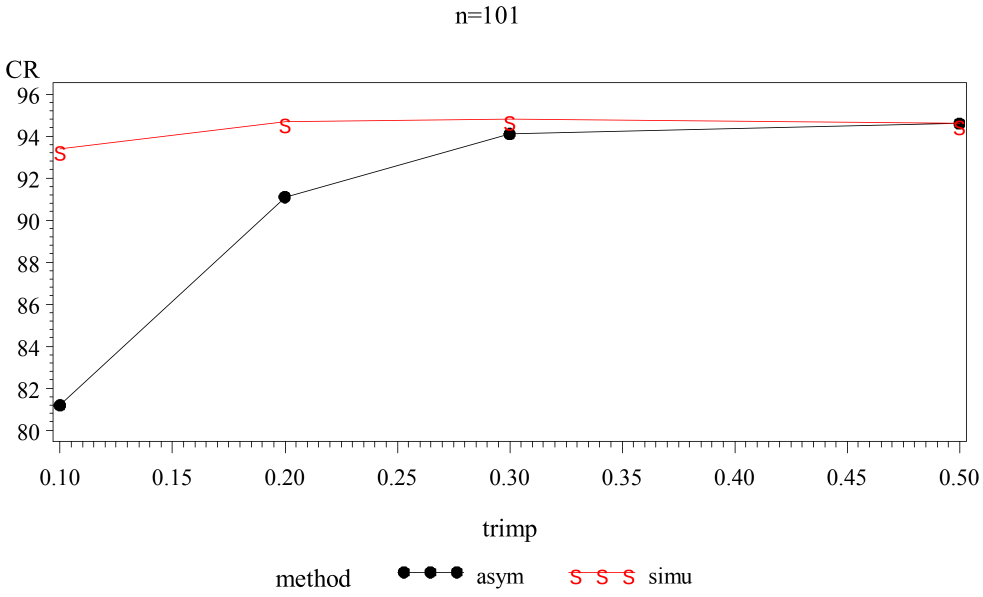
b) *n*=21



c) *n*=31


n=31
CR
96
94
92
90
88
86
84
82
80
78
76
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
trimp
method
asym
S S S simu

d) *n*=101


n=101
CR
96
94
92
90
88
86
84
82
80
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
trimp
method
asym
S S S simu

The plots of *CR* versus *n* with respect to the two methods at *p*=0.1, 0.2, 0.3 and 0.5 are given in Figure 4.2 a)-d) respectively. Non-parallelism was observed in the plots, which supports the statistical significance of the effect between *n* and method, in the regression analysis in Table 4.4. In general, CR for method II is higher than those of method I. Figure 4.3a) and b) are the plots of *CR* versus trimming proportion *p* for the two methods with respect to *n*. Non-parallelism was also observed, which supports the statistical significance of the effect between *p* and method in the regression analysis in Table 4.4. It seems that for Method I, the coverage rate goes up as the trimming proportion *p* increases, while for Method II method, *p* does not have a big effect on coverage rate. Graphs of *CR* versus *p* with respect to Method I and II at *n*=11, 21, 31, 101 are presented in Figure 4.4 a)-d), respectively. Relative parallelism was observed in the *n* by *p* plots, which is consistent with the non- statistical significance of the interaction between *n* and *p* in Table 4.4. From Figures 4.4, the relative parallelism of the profiles hints that any interaction that may exist between *n* and *p* is not of practical importance. It appears that at each specific sample size n, as *p* increases, for Method I, *CR* increases, while for Method II,*CR* does not change very much, which further indicates that the differences between the coverage rates of the methods depend on trimming rate *p*.

Overall, these plots support the conclusions that in terms of coverage rate: (i) Method II is better and more stable across conditions than Method I; (ii) Method I improves as *n* increases, except for *n* = 21; (iii) Method I improves as *p* increases from 0.10 to 0.30 and then levels off. The later observation may be due to the fact that increasing trimming discards outliers up to a point where the remaining values are relatively well behaved.

4.6 Assessments of Average Length Method I and II Cauchy Data

The average lengths of the confidence intervals in tables 4.1 and 4.2 were also compared for Method I and II in terms of sample size *n* and trimming rate *p*. Proc GPLOT was again used to make plots to visualize the average lengths in terms of *n*, *p* and Method. Proc GLM was used to fit a linear model with average interval length as the response and sample size *n*, method and trimming proportion *p* as independent variables, as given in section 4.5.

The output from Proc GLM for average length as response is given in Table 4.5 below. The coefficient of determination for the model $R^2 = 0.86$ and the plot of residuals and the normality QQ plot are given in Figure 4.5a) and b). Although these plots indicate possible shortcomings of the model and raise questions about the assumption of normality, they can still be useful as guides to identifying significant sources of variation. The only statistically significant effect is sample size *n*, which has estimated regression coefficient equal to negative 0.0135. Thus, all other effects being held fixed, we estimate that for either Method I or Method II, average confidence interval length decreases by 0.0135 per unit increase in sample size.

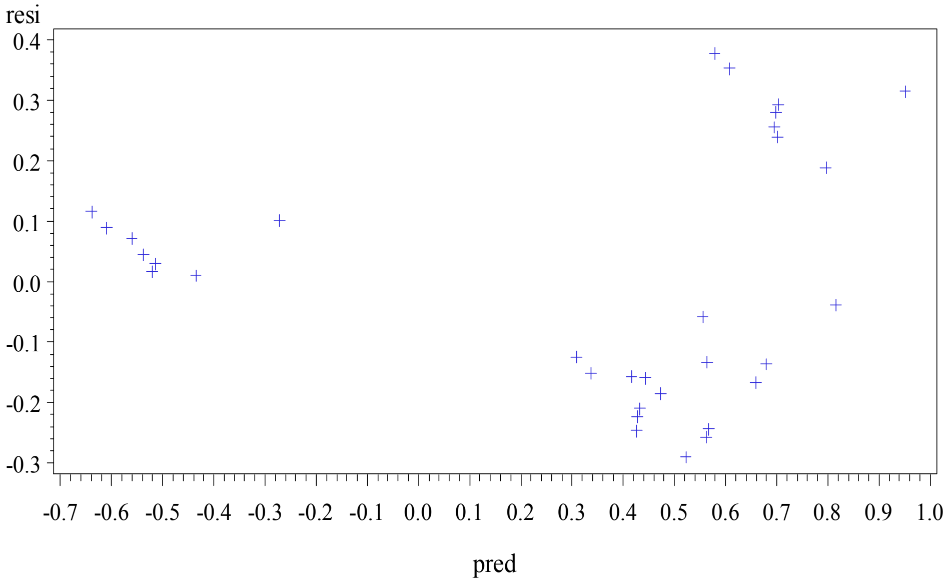
Table 4.5 Output of Proc GLM using Average Length as Response for Cauchy data

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	7.81419709	0.97677464	17.47	<.0001
Error	23	1.28584528	0.05590632		
Corrected Total	31	9.10004237			

R-Square	Coeff Var	Root MSE	aveCIW2 Mean
0.858699	76.90711	0.236445	0.307443

Source	DF	Type III SS	Mean Square	F Value	Pr > F
N	1	1.64160370	1.64160370	29.36	<.0001
method	1	0.07738558	0.07738558	1.38	0.2514
trimp	1	0.11110120	0.11110120	1.99	0.1720
n*method	1	0.00000439	0.00000439	0.00	0.9930
n*trimp	1	0.00015492	0.00015492	0.00	0.9585
trimp*method	1	0.03075777	0.03075777	0.55	0.4658
n*trimp*method	1	0.00019454	0.00019454	0.00	0.9535
trimp*trimp	1	0.09262202	0.09262202	1.66	0.2108

Figure 4.5 a)   Residuals Plot Average Length Cauchy Data



b)Normality QQ Plot Average Length Cauchy Data

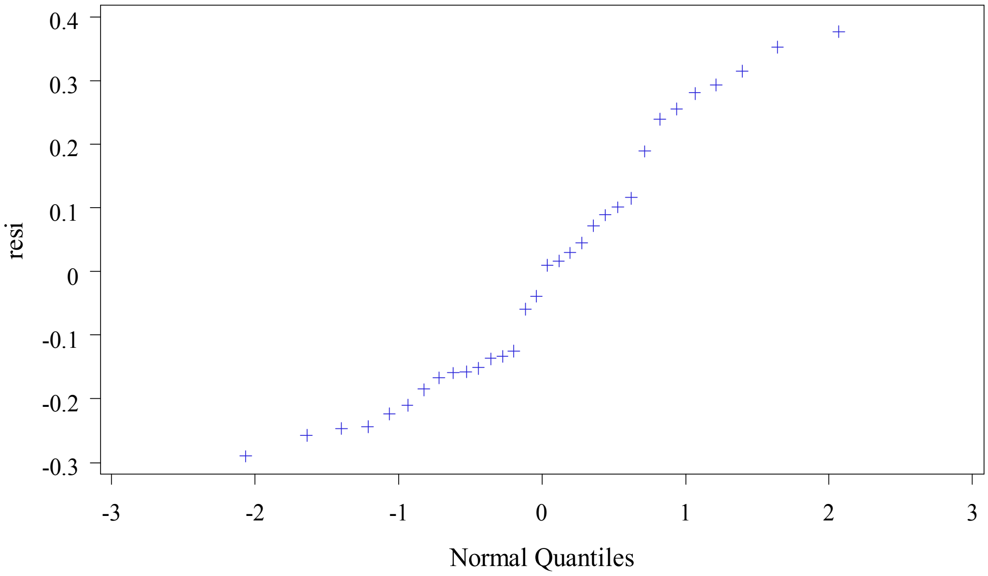
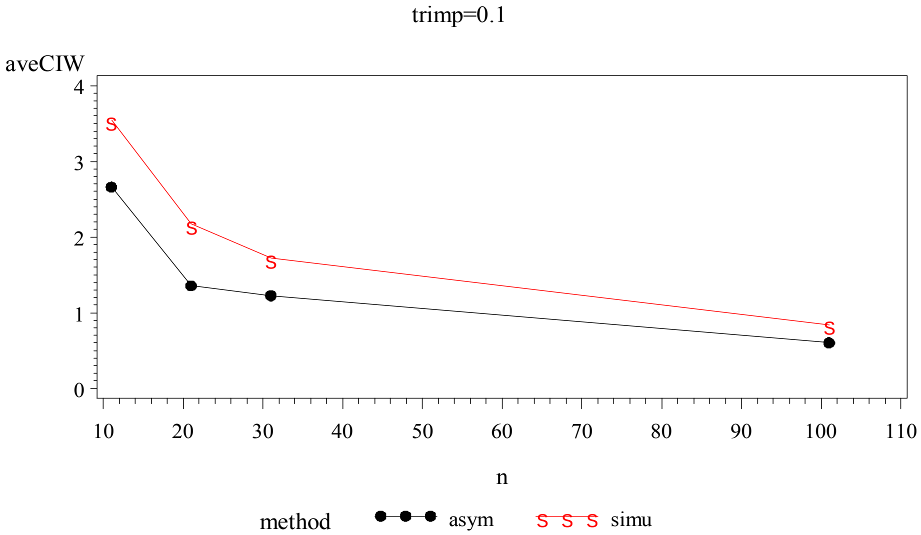
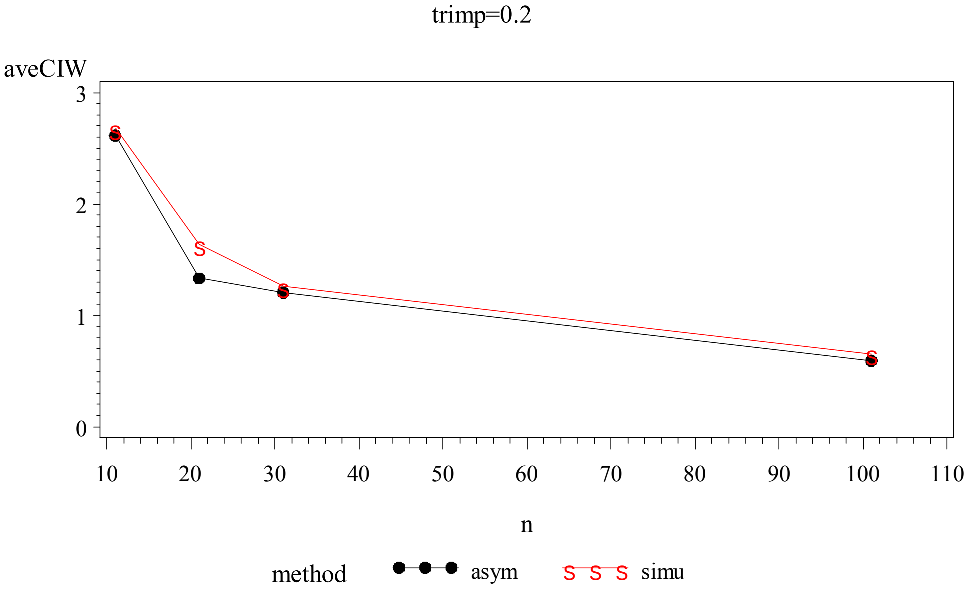


Figure 4.6 Average Length: Sample Size by Methods I and II Interaction Cauchy Data

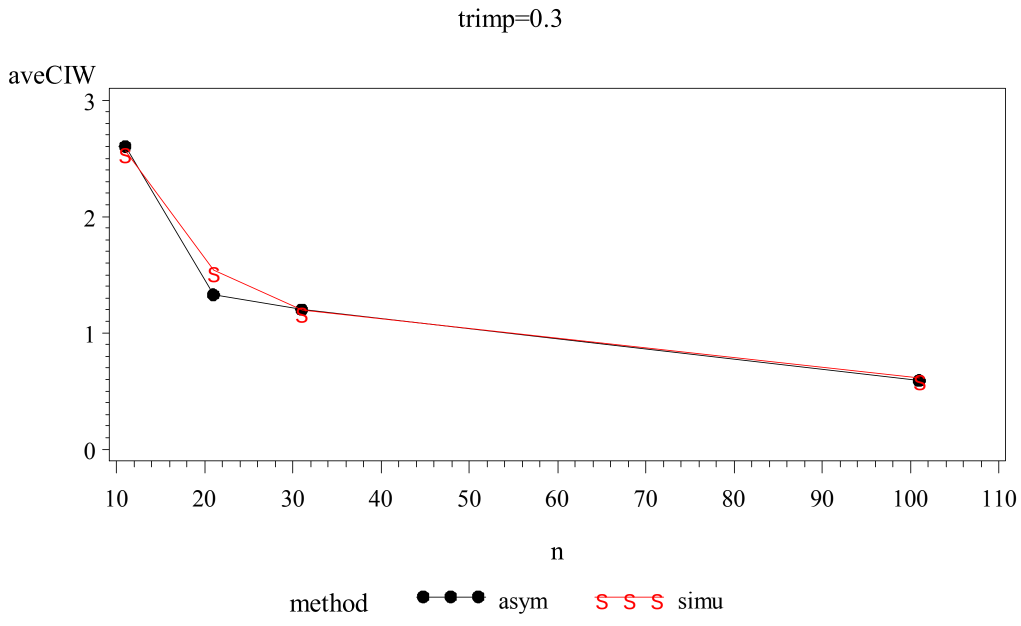
a) *p* =0.1



b) *p*=0.2



c) *p*=0.3


trimp=0.3
aveCIW
3
2
1
0
10
20
30
40
50
60
70
80
90
100
110
n
method
asym
S S S
simu

d) *p*=0.5

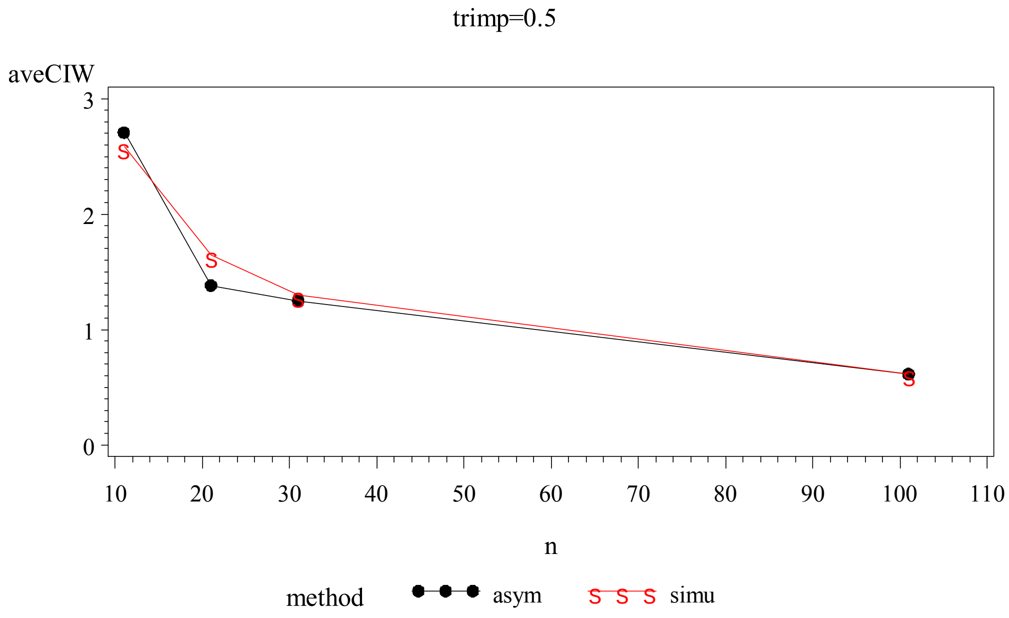
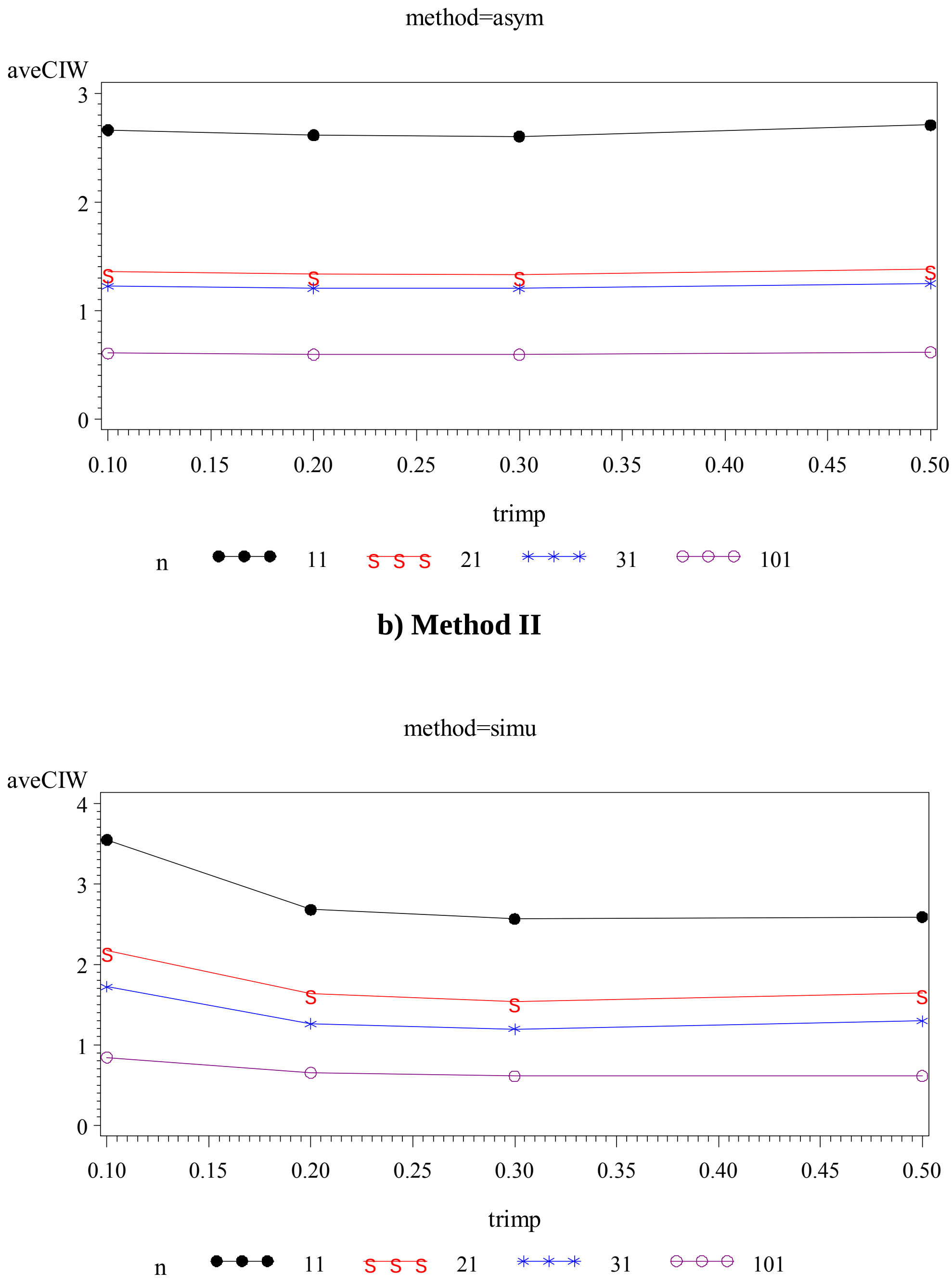

trimp=0.5
aveCIW
3
2
1
0
10
20
30
40
50
60
70
80
90
100
110
n
method
asym
S S S
simu

Figure 4.7 Average Length: Sample Size by Trimming Proportion Interaction

Cauchy Data

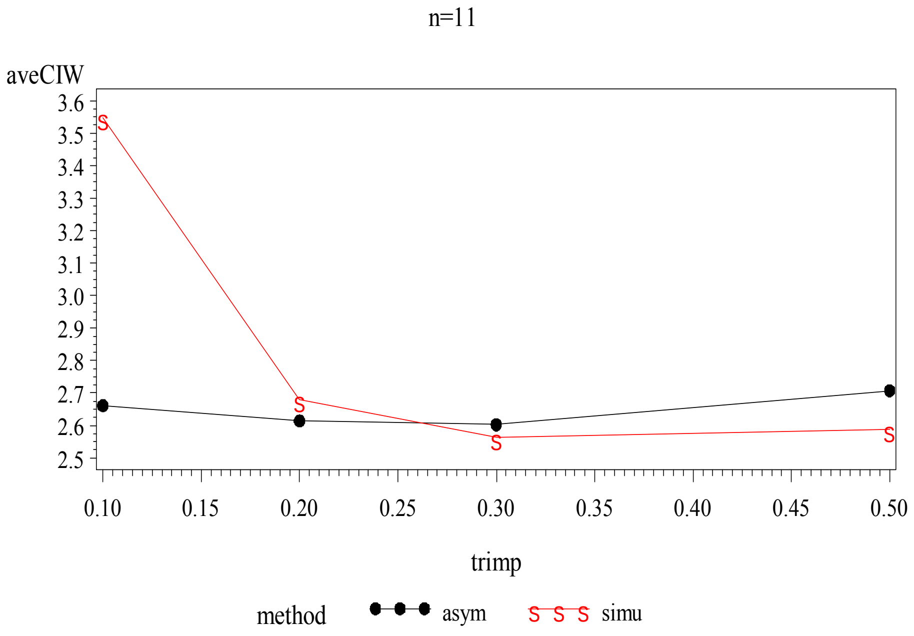
a) Method I



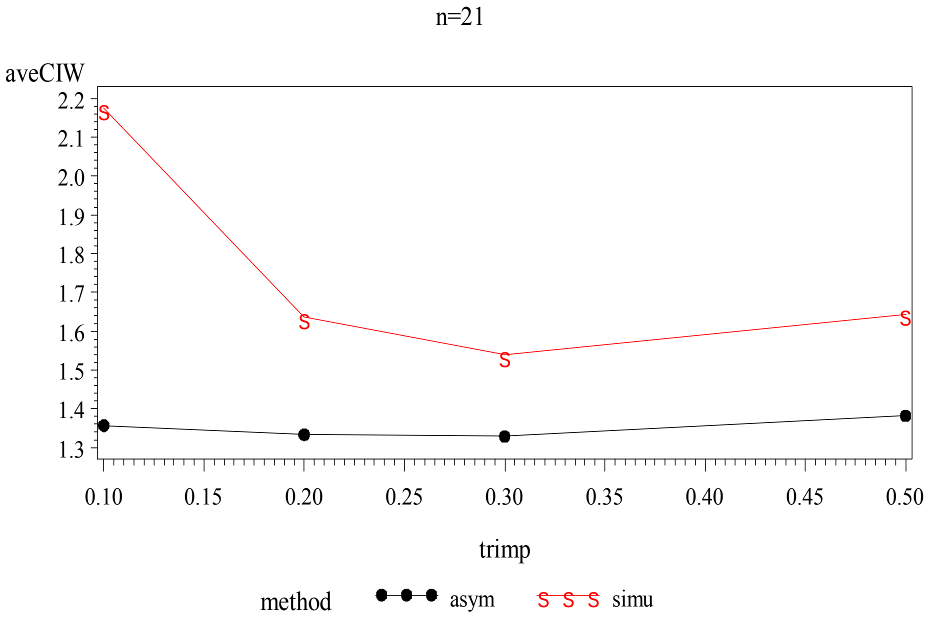
b) Method II

Figure 4.8 Average Length: Trimming Proportion by Method Interaction Cauchy Data

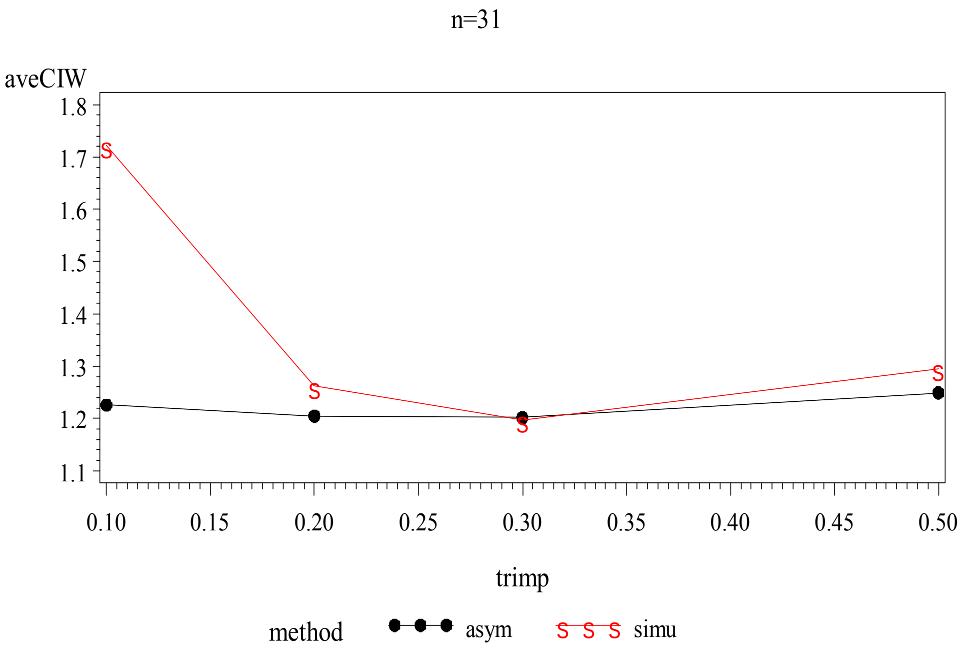
a) *n*=11



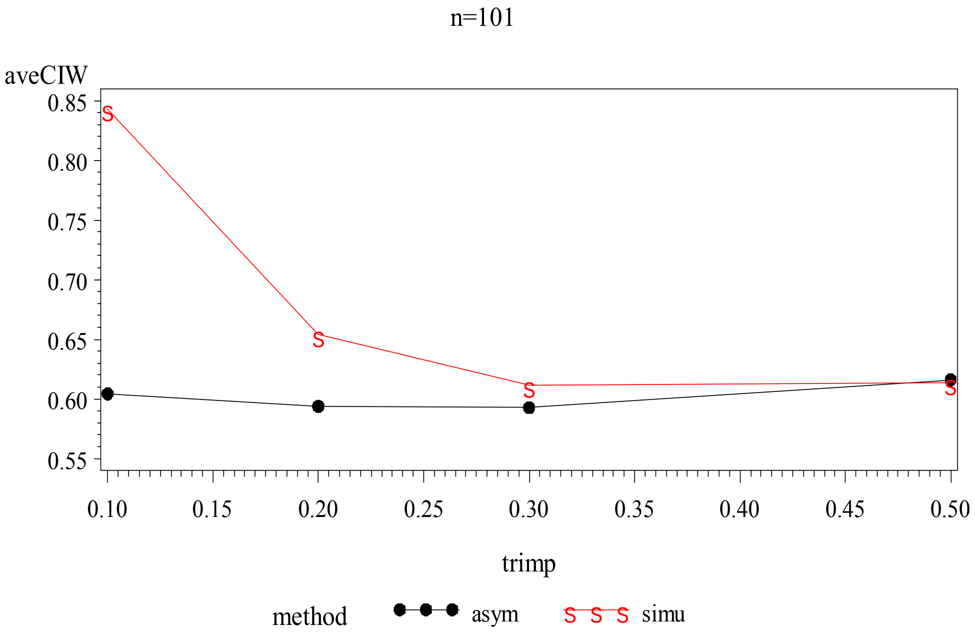
b) *n*=21



c) *n*=31


n=31
aveCIW
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
trimp
method
asym
simu

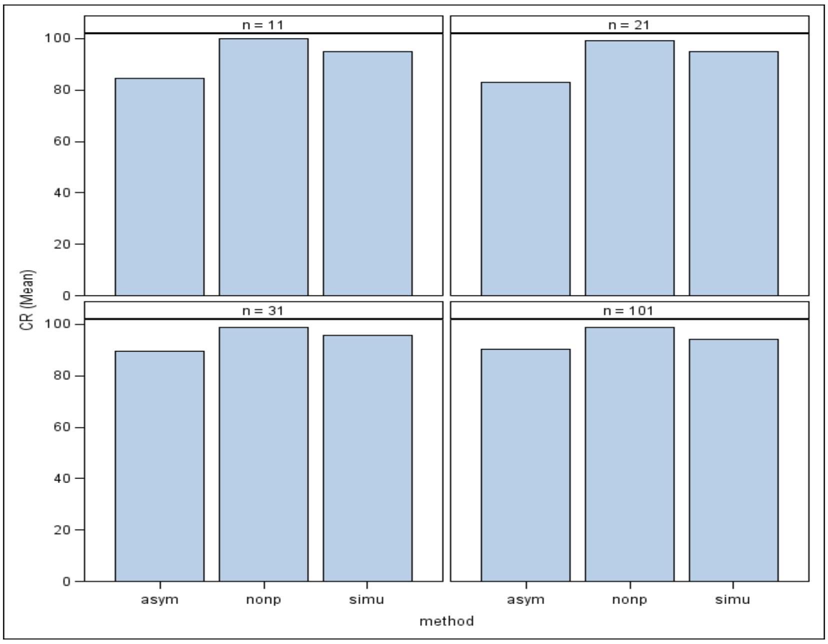
d) *n*=101


n=101
aveCIW
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
trimp
method
asym
simu

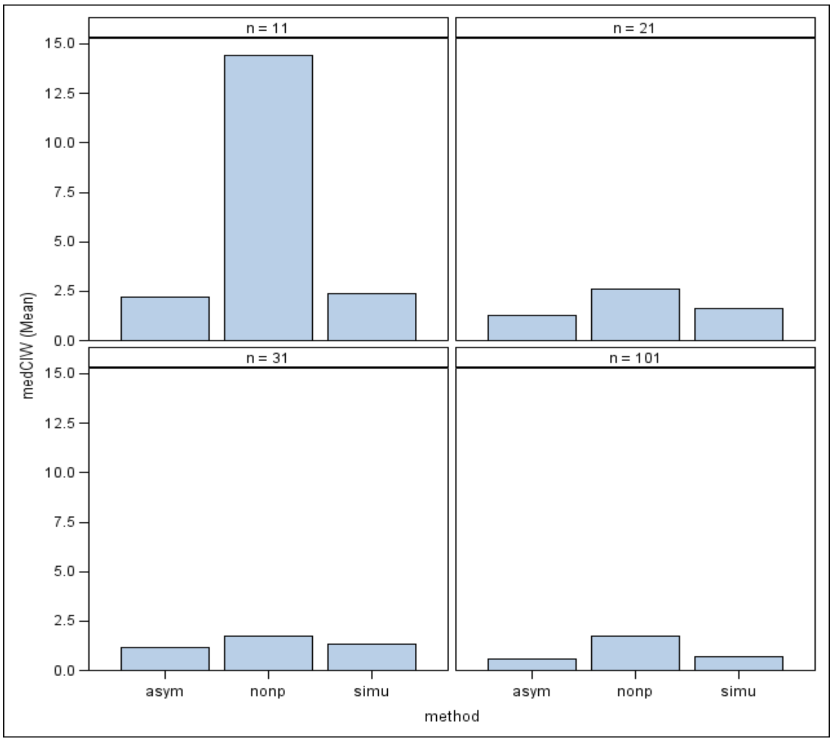
The graphs of average lengths versus sample size *n* at *p*=0.1, 0.2, 0.3 and 0.5 with respect to the two methods are displayed in Figure 4.6 a)-d), respectively. It appears, as expected, that average length decreases as *n* increases for all four graphs. Similar trends were observed for the two methods and for the four trimming proportions. Figure 4.7 a) and b) are the plots of average lengths versus trimming proportion *p* for the two methods with respect to *n*. It seems that for the two methods, the average lengths do not change much as *p* changes. Similar approximate parallelism was observed for both of the graphs, indicating there was no interaction between p and method, between *n* and *p*. Graphs of average lengths versus trimming proportion *p* with respect to methods at sample sizes *n*=11, 21, 31, 101 are presented in Figures 4.8 a)-d), respectively. It appears that at each specific sample size, as trimming proportion *p* increases, for Method I, average lengths do not change much, while for Method II, average length decrease in general. Overall, average length for Method II is greater than that for Method I, which may explain why the coverage rates for Method I tend to be lower than nominal, especially for small *n*. Also, the average length for Method II for the smallest trimming *p* = 0.1 is considerably greater than that for *p* > 0.1. This may be due to the fact that the data still contains large outliers for *p* =0 .1, which inflates trimmed means and simulated estimates of the quantiles of the statistics $t_{\alpha/2,n}, t_{1-\alpha/2,n}$, given in Tables 4.1-4.2.

For Method III, the nonparametric method, coverage rates tend to be a bit above their nominal 0.95 value and also greater than those from Method I and II, as given in Figure 4.9a). Correspondingly median lengths of Method III intervals, which decrease with increasing sample size *n*, are greater than those from Method I and II, as illustrated in Figure 4.9b). Median length was used instead here because the average length for *n*=11 is distorted by outliers.

Figure 4.9 a) Average of CR across *p* against Method by *n*



b) Median of CI lengths across p against Method by n



Chapter 5 Robustness Study

I also briefly investigated the performance of the confidence intervals constructed above assuming the Cauchy model when the data are actually sampled from the normal and logistic distributions, two widely used symmetric distributions. Thus, the quantiles estimated in Chapter 4 were used here to construct intervals under the erroneous belief that the data are sampled from a Cauchy distribution. In order to facilitate comparison of interval length, I scaled both distributions so that they have interquartile ranges (IQR) equal to 2.0, the IQR of the standard Cauchy distribution, as follows:

Normal Data: Generate $Y \sim N(0,1)$ and let X = (2/1.35)Y (9)

Logistic Data: Generate $Y \sim Logistic(0,1)$ and let X = (2/2.20)Y (10)

To carry out my study, I used the same settings as for Cauchy data. Specifically, I set nominal coverage rate $1-\alpha$ = 0.95 and used the same representative values of *n* = 2*m*+ 1 and *p*. In the simulation algorithm given below, 'D' stands for either the logistic or normal distribution.

(1) Generate R = 1000 independent random samples $\{z_i^* = (z_{i1}, z_{i2}, \ldots, z_{in})\}$, *i*= 1,2,…, *R*, from the standard Cauchy distribution. I actually used the same data generated in Chapter 4.

(2) Generate a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from distribution D.

(3) Let $\hat{\sigma}_3$ equal the sample semi-interquartile range, as defined in (6), computed from $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

(4) For Method I, use the formulas $t_{\alpha/2,n} = -\gamma z_{\alpha/2} / \sqrt{(nk)}$ and $t_{1-\alpha/2,n} = \gamma z_{\alpha/2} / \sqrt{nk}$, resulting in the same values as those obtained in Chapter 4.

(5) For Method II, use the values $\hat{t}_{\alpha/2,n} = \hat{H}_n^{-1}(\alpha/2)$ and $\hat{t}_{1-\alpha/2,n} = \hat{H}_n^{-1}(1-\alpha/2)$ obtained in Chapter 4 from Cauchy data.

(6) Construct confidence intervals using all three methods. Record whether or not each interval contains $\mu = 0$ and its length.

(7) Independently repeat steps of (2)-(6) N =1000 times.

Assess and compare the performance of the confidence intervals across all parameter settings.

5.1 Normal Data Results

Table 5.1 Simulated Coverage Rates, Average Lengths Method I Normal Data

		P=0.1	P=0.2	P=0.3	P=0.5
n=11	Estimated t	0.912	0.896	0.892	N/A
	Coverage Rate	90.2	90	89.7	90.9
	Average CI Length	1.98	1.95	1.94	2.01
	Median of CI Length	1.91	1.88	1.87	1.94
n=21	Estimated t	0.659	0.648	0.646	N/A
	Coverage Rate	87.8	86.2	86.6	86.5
	Average CI Length	1.22	1.2	1.195	1.24
	Median of CI Length	1.206	1.185	1.18	1.23
n=31	Estimated t	0.542	0.533	0.532	N/A
	Coverage Rate	92.2	91.5	91.5	91
	Average CI Length	1.11	1.093	1.09	1.133
	Median of CI Length	1.1	1.082	1.08	1.22
n=101	Estimated t	0.3	0.295	0.2946	N/A
	Coverage Rate	93.8	92.4	90.9	89.4
	Average CI Length	0.593	0.583	0.582	0.605
	Median of CI Length	0.591	0.581	0.58	0.603

Table 5.2 Estimated Coverage Rates, Average Lengths Based on Method II Normal Data

		P=0.1	P=0.2	P=0.3	P=0.5
n=11	Estimated t (Cauchy Result)	1.216	0.918	0.878	0.887
	Coverage Rate	97.6	92.7	90.5	89.1
	Average CI Length	2.64	1.99	1.907	1.93
	Median of CI Length	2.55	1.92	1.84	1.86
n=21	Estimated t(Cauchy Result)	1.056	0.795	0.748	0.799
	Coverage Rate	98.2	92.4	91.2	92
	Average CI Length	1.96	1.47	1.384	1.48
	Median of CI Length	1.94	1.45	1.368	1.46
n=31	Estimated t(Cauchy Result)	0.762	0.558	0.53	0.573
	Coverage Rate	99.2	93.9	91.8	92.1
	Average CI Length	1.56	1.14	1.087	1.175
	Median of CI Length	1.55	1.13	1.076	1.163
n=101	Estimated t(Cauchy Result)	0.419	0.325	0.304	0.305
	Coverage Rate	99.4	95.1	92.2	89.3
	Average CI Length	0.827	0.641	0.6	0.6
	Median of CI Length	0.825	0.64	0.599	0.6

Table 5.3 Estimated Coverage Rates, Average Lengths Based on Method III Normal data

	n=11	n=21	n=31	n=101
Coverage Rate	100	99	98.9	97
Average CI Length	4.71	2.31	1.762	0.808
Median of CI Length	4.66	2.28	1.748	0.799

5.2 Logistic Data Results

Table 5.4. Estimated Coverage Rates, Average Lengths Method I Logistic Data

		P=0.1	P=0.2	P=0.3	P=0.5
n=11	Estimated t	0.912	0.896	0.892	N/A
	Coverage Rate	88.2	89.4	90.7	90.7
	Average CI Length	2.02	1.98	1.98	2.05
	Median of CI Length	1.98	1.94	1.93	2.01
n=21	Estimated t	0.659	0.648	0.646	N/A
	Coverage Rate	87.3	87.8	87	86.1
	Average CI Length	1.24	1.22	1.215	1.26
	Median of CI Length	1.22	1.2	1.195	1.24
n=31	Estimated t	0.542	0.533	0.532	N/A
	Coverage Rate	92.2	91.8	91.2	90.1
	Average CI Length	1.13	1.11	1.108	1.15
	Median of CI Length	1.123	1.1	1.1	1.14
n=101	Estimated t	0.3	0.295	0.2946	N/A
	Coverage Rate	92.8	92.8	92	90.5
	Average CI Length	0.592	0.582	0.581	0.604
	Median of CI Length	0.591	0.581	0.58	0.603

Table 5.5. Estimated Coverage Rates, Average Lengths Based on Method II Logistic Data

		P=0.1	P=0.2	P=0.3	P=0.5
n=11	Estimated t (Cauchy Result)	1.216	0.918	0.878	0.887
	Coverage Rate	96.6	92.1	90.4	89.8
	Average CI Length	2.69	2.03	1.94	1.96
	Median of CI Length	2.64	1.99	1.9	1.92
n=21	Estimated t(Cauchy Result)	1.056	0.795	0.748	0.799
	Coverage Rate	98.8	94.3	92	92.2
	Average CI Length	1.99	1.5	1.41	1.5
	Median of CI Length	1.96	1.47	1.39	1.48
n=31	Estimated t(Cauchy Result)	0.762	0.558	0.53	0.573
	Coverage Rate	98.4	93.7	91.8	91.1
	Average CI Length	1.59	1.63	1.105	1.2
	Median of CI Length	1.56	1.16	1.1	1.19
n=101	Estimated t(Cauchy Result)	0.419	0.325	0.304	0.305
	Coverage Rate	92.8	92.8	92	90.5
	Average CI Length	0.592	0.582	0.581	0.604
	Median of CI Length	0.591	0.581	0.58	0.603

Table 5.6 Estimated Coverage Rates, Average Lengths Method III Logistic Data

	n=11	n=21	n=31	n=101
Coverage Rate	99.8	99.7	98.9	97.4
Average CI Length	5.35	2.37	1.779	0.808
Median of CI Length	5.13	2.36	1.744	0.797

5.3 Comparison of CR and Average length Cauchy, Normal and Logistic Data Method II

The plots of average of CR and average of average length across *p* against distribution at *n*=11, 21, 31 and 101 are displayed in Figure 5.1 and 5.2 respectively. In Figure 5.1, it can be seen that the coverage rates for the three distribution data are similar, with a slight but noticeable decrease as trimming proportion increases for intervals constructed from normal and logistic data (see table5.2 and table 5.5). This may happen because increasing trimming removes ‘information’ from these relatively light tailed distributions. In Figure 5.2, for *n*=11 the average of average CI lengths from Cauchy data are wider than those of Normal and Logistic data. But, the averages of average lengths are quite similar among the three distributions for the larger sample sizes. As expected, the averages of average lengths decrease as sample size *n* increases. We could conclude that Method I and II are reasonably robust with respect to departures from an assumed Cauchy model in the settings I studied.

Figure 5.1 Average of Coverage Rate across *p* vs Distribution by *n*

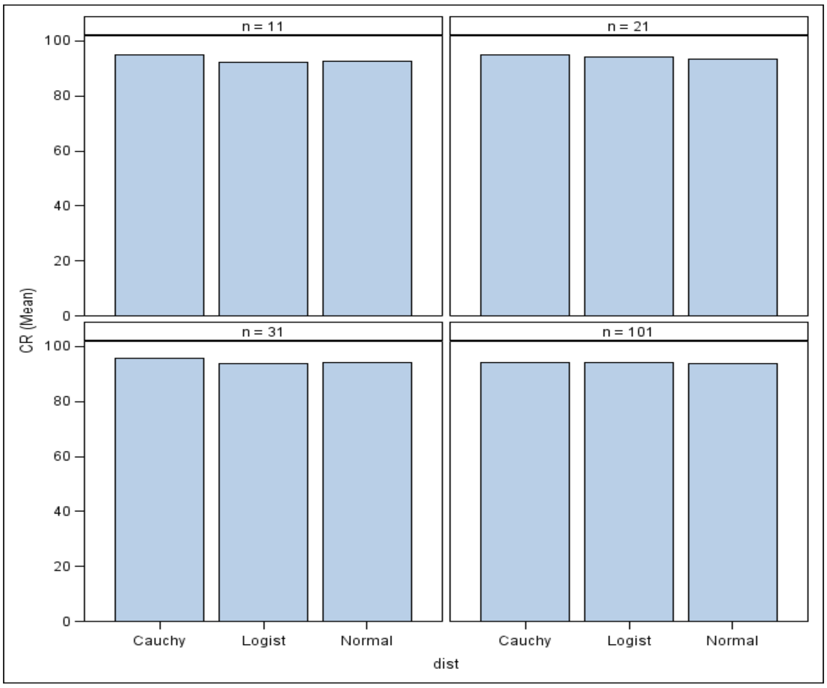
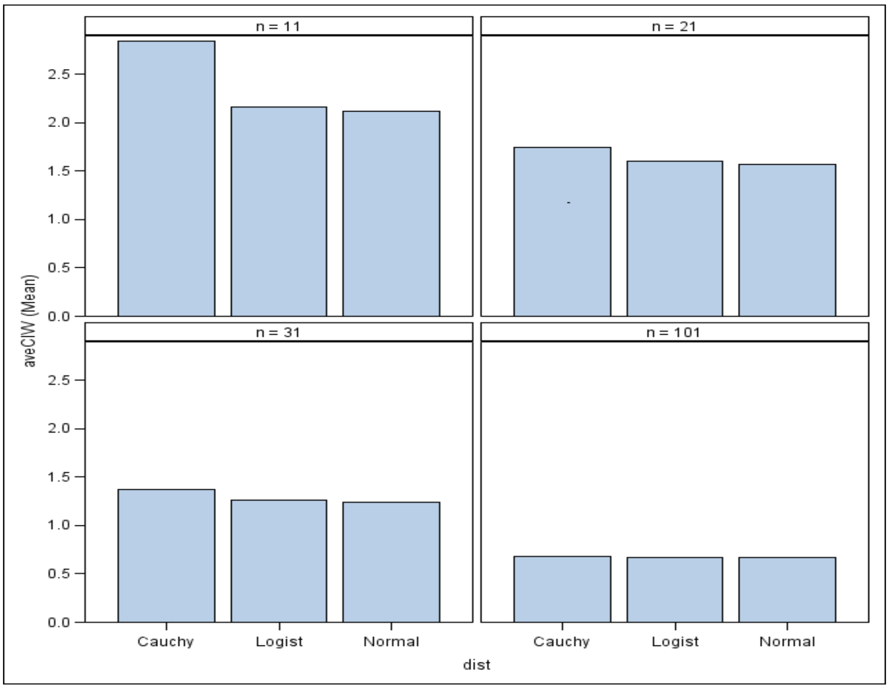


Figure 5.2 Average of Average lengths across *p* vs Distribution by *n*



Chapter 6 Summaries

The Cauchy distribution is important as an example of a pathological case. Cauchy distributions look similar to a normal distribution. However, they have much heavier tails. When studying hypothesis tests that assume normality, seeing how the tests and confidence intervals perform on data from a Cauchy distribution is a good indicator of how sensitive the tests are to heavy-tail departures from normality. Likewise, using it is a good check for robust techniques that are designed to work well under a wide variety of distributional assumptions.

This report investigated the performance of confidence intervals for the median of a Cauchy distribution based on trimmed means when the data are sampled from Cauchy, normal and logistic distributions. Actual coverage rate of nominal 0.95 confidence intervals and their average lengths were used as criteria for judging performance. Two methods, as described in Chapter 2.3, were used to estimate the quintiles of exact 0.95 equivariant confidence intervals for the median. A nonparametric method based on order statistics was also used to construct confidence intervals, as discussed in Chapter 3. My simulation study led to the following conclusions.

1) Coverage Rate Cauchy Data:

In terms of coverage rate, (i) Method II is better and more stable across all conditions than Method I; (ii) Method I improves as sample size n increases, except for $n = 21$; (iii) Method I improves as trimming proportion p increases from 0.10 to 0.30 and then levels off. The later observation may be due to the fact that increasing trimming discards outliers up to a point where the remaining values are relatively well behaved. For Method III, coverage rates tend to be a bit above their nominal 0.95 value and also greater than those from Method I and II.

2) Average Length Cauchy Data:

For Method I, the average lengths do not change much as p changes, while for Method II, the average lengths decrease in general as p increases. Overall, the average lengths for Method II are greater than those for Method I, which may explain why the coverage rates for Method I tend to be lower than nominal, especially for small n. Also, the average lengths for Method II for the smallest trimming, $p = 0.1$ is considerably greater than that for $p > 0.1$. This may be due to the fact that the data still contains large outliers for p =0 .1, which inflates trimmed means and simulated estimates of the quantiles of the statistics $t_{\alpha/2,n}, t_{1-\alpha/2,n}$.

The average lengths of Method III intervals are greater than those from Method I and II.

3) Robustness Study:

It was observed that the coverage rates for data from the three distributions are similar. At sample size $n = 11$, the average CI lengths from Cauchy data are wider than those of normal and logistic data, and are quite close among the three distributions for the larger sample sizes.

4) Overall Summary:

Method II, based on using simulation to estimate the quantiles of the exact, equivariant intervals performed better than the other two methods for all three distributions in the settings I studied. I recommend using this method with p= 0.2 and n >30 to estimate the median of a Cauchy distribution. It also performed reasonably well when data were actually sampled from normal and logistic distributions.

Bibliography

[1] Chen, Z. (2011): ‘A Simple Method for Estimating Parameters of the Location-Scale Distribution Family.’ *Journal of Statistical Computation and Simulation*, 81, 49-58.

[2] David, H.A. (1970): Order Statistics. Wiley,NYC.

[3] T.J., Fisher, F.M., Tilanus, C.B. (1964): ‘A Note on Estimation From a Cauchy Sample’. *JASA*, *59*, 460-463.

[4] Stigler, S.M. (1973): ‘The Asymptotic Distribution of the Trimmed Mean’, *Ann. Statist.*, *1*, 472-477.

[5] Serfling, R.(2001), ‘Approximation Theorems of Mathematical Statistics’, Wiley,NYC

[6] Hogg, R.(2004), ‘Introduction to Mathematical Statistics’, Pearson Education. NJ

[7] Wilcox, R., Keselman, H.J. ‘Repeated measures one-way Anova based on a modified one-step M-estimator British Journal of Mathematical&Statistical Psychology’, 2003, 56,15-25.

[8] Yang, J., ‘Outlier identification and robust parameter estimation in a zero-inflated Poisson model’, Journal of Applied Statistics, 2011,38, 421-430.

[9] Walfish, S. A review of Statistical Outlier Methods. Pharmaceutical Technology. 2,2006.

[10] Jacob, J. and Protter P.(1998) Probability Essentials

[11] Furguson,T.(1996) A Course in Large Sample Theory, CRC Press.

Appendix A -

SAS code for Asymptotic Normality Method

```
ods rtf file="t:/asymp.rtf";
ods listing close;
libname ms "t:/";
%macro msreportv(seed, repeat,n,r);

%do j=1 %to &repeat;
data cau_&j(drop=i);
%Do i =1 %to &n;
w=rancau(&seed+&j);
output;
%end;
run;

proc sort data=cau_&j;
by w;
run;
proc means data=cau_&j noprint;
var w;
output out=a_&j p25=q1 p75=q3;

data caunew_&j;
set cau_&j;
xx=_n_;
if  xx >= &n-&r+2 or xx<=&r then delete;
proc means data=caunew_&j noprint;
var w;
output out=s_&j mean=mu;
run;
data as_&j;
merge a_&j s_&j;
run;
data as1_&j;
set as_&j;
stdev=0.5*(q3-q1);


run;

%end;

%mend
;

%msreportv(789569,1000,101,30)
```

```
%macro namesv(prefix,maxnum);
%do i=1 %to &maxnum;
&prefix&i
%end;
;

%mend namesv;

data alldiff;
set %namesv(as1_,1000);
run;
%let a=101;
%let b=30;
%let m=62;

data ms.alldd&m;
set alldiff;
p=&b/&a;
k=1/2-p;
f=sqrt((1-
k)*tan(3.14159*k/2)*tan(3.14159*k/2)/k+2*tan(3.14159*k/2)/(3.14159*k)-1);
t=f*1.96/sqrt(&a*k);
LLCI=mu-t*stdev;
ULCI=mu+t*stdev;
CIW&m=ULCI-LLCI;

data ms.alldiffd&m(keep=CIW&m c&m);
set ms.alldd&m;
if LLCI <=0 and ULCI >=0 then c&m=1;
else c&m=0;
data dd1;
set ms.alldiffd&m(drop=CIW&m);
proc transpose data=dd1 out=transc;
data transc1;
set transc;
CR=sum(of col1-col1000)/1000*100;
run;
data dd2;
set ms.alldiffd&m(drop=c&m);
proc transpose data=dd2 out=transciw;
data transciw1;
set transciw;
aveCIW=mean(of col1-col1000);
medCIW=median(of col1-col1000);
data ms.ciwc&m;
merge transc1(drop=_name_ col1-col1000) transciw1(drop=_name_ col1-col1000);
proc print data=ms.alldiffd&m;
proc print data=ms.ciwc&m;
proc print data=ms.alldd&m;

run;

ods rtf close;
ods listing;
```

```
/* code for sample median;
libname ms "t:/";
%macro msreportv(seed, repeat,n,r);

%do j=1 %to &repeat;
data cau_&j(drop=i);
%Do i =1 %to &n;
w=rancau(&seed+&j);
output;
%end;
run;

proc sort data=cau_&j;
by w;
run;
proc means data=cau_&j noprint;
var w;
output out=a_&j p25=q1 p75=q3;

data caunew_&j;
set cau_&j;
xx=_n_;
if  xx >= &n-&r+1 or xx<=&r then delete;
proc means data=caunew_&j noprint;
var w;
output out=s_&j mean=mu;
run;
data as_&j;
merge a_&j s_&j;
run;
data as1_&j;
set as_&j;
stdev=0.5*(q3-q1);


run;

%end;

%mend
;

%msreportv(789569,1000,101,50)


%macro namesv(prefix,maxnum);
%do i=1 %to &maxnum;
&prefix&i
%end;
;

%mend namesv;

/* Call the macro on the SET statement */

data alldiff;
```

```
set %namesv(as1_,1000);
run;
%let a=101;
%let m=63;

data ms.alldd&m;
set alldiff;
LLCI=mu-1.96*3.14*stdev/(2*sqrt(&a));
ULCI=mu+1.96*3.14*stdev/(2*sqrt(&a));
CIW&m=ULCI-LLCI;

data ms.alldiffd&m(keep=CIW&m c&m);
set ms.alldd&m;
if LLCI <=0 and ULCI >=0 then c&m=1;
else c&m=0;
/*proc sort data=ms.alldiffd&m;
by CIW&m;*/
data dd1;
set ms.alldiffd&m(drop=CIW&m);
proc transpose data=dd1 out=transc;
data transc1;
set transc;
CR=sum(of col1-col1000)/1000*100;
run;
data dd2;
set ms.alldiffd&m(drop=c&m);
proc transpose data=dd2 out=transciw;
data transciw1;
set transciw;
aveCIW=mean(of col1-col1000);
medCIW=median(of col1-col1000);
data ms.ciwc&m;
merge transc1(drop=_name_ col1-col1000) transciw1(drop=_name_ col1-col1000);
proc print data=ms.alldiffd&m;
proc print data=ms.ciwc&m;
proc print data=ms.alldd;


run;
ods rtf close;
ods listing;
```

Appendix B - SAS code for Simulation Method

```
ods rtf file="t:/simu.rtf";
ods listing close;
options nonotes nosource nosource2 errors=0;
libname ms "t:/";
%macro msreport(seed, repeat, n, r);

%do j=1 %to &repeat;
data cau_&j(drop=i);
%Do i =1 %to &n;
w=rancau(&seed+&j);
 output;
%end;
run;
proc sort data=cau_&j;
by w;
run;
proc means data=cau_&j noprint;
var w;
output out=a_&j p25=q1 p75=q3;


data caunew_&j;
set cau_&j;
xx=_n_;
if  xx >= &n-&r+1 or xx<=&r then delete;
proc means data=caunew_&j noprint;
var w;
output out=s_&j mean=mu;
run;
data as_&j;
merge a_&j s_&j;
run;
data as1_&j;
set as_&j;
stdev=0.5*(q3-q1);
T=mu/stdev;

*proc print data=as1_&j;
run;
%end;

%mend
;

%msreport(12345,1000,31,15)

%macro names(prefix,maxnum);
%do i=1 %to &maxnum;
&prefix&i
%end;
```

```
;

%mend names;

data all;
set %names(as1_,1000);
run;
proc sort data=all;
by T;
data all1;
set all;
nn=_n_;
 if nn=25 then LL=T;
 If nn=975 then UL=T;

 proc sql;
 select LL into: LLC
 from all1
 where LL ne .;
 select UL into:ULC
 from all1
 where UL ne .;
quit;

%macro msreportv(seed, repeat,a,b); *a=n, b=r;

%do j=1 %to &repeat;
data cau_&j(drop=i);
%Do i =1 %to &a;
w=rancau(&seed+&j);
output;
%end;
run;

proc sort data=cau_&j;
by w;
run;
proc means data=cau_&j noprint;
var w;
output out=a_&j p25=q1 p75=q3;

data caunew_&j;
set cau_&j;
xx=_n_;
if  xx >= &a-&b+1 or xx<=&b then delete;
proc means data=caunew_&j noprint;
var w;
output out=s_&j mean=mu;
run;
data as_&j;
merge a_&j s_&j;
run;
data as1_&j;
set as_&j;
stdev=0.5*(q3-q1);
```

```
%end;

%mend
;

%msreportv(789569,1000,31,15)
%macro namesv(prefix,maxnum);
%do i=1 %to &maxnum;
&prefix&i
%end;
;

%mend namesv;


data alldiff;
set %namesv(as1_,1000);
run;
%let m=47;
data ms.alldd&m;
set alldiff;
t2=(abs(&LLC)+abs(&ULC))/2;
LLCI=mu-t2*stdev;
ULCI=mu+t2*stdev;
CIW&m=ULCI-LLCI;


data ms.alldiffd&m(keep=CIW&m c&m);
set ms.alldd&m;
if LLCI <=0 and ULCI >=0 then c&m=1;
else c&m=0;
data dd1;
set ms.alldiffd&m(drop=CIW&m);
proc transpose data=dd1 out=transc;
data transc1;
set transc;
CR=sum(of col1-col1000)/1000*100;
run;
data dd2;
set ms.alldiffd&m(drop=c&m);
proc transpose data=dd2 out=transciw;
data transciw1;
set transciw;
aveCIW=mean(of col1-col1000);
medCIW=median(of col1-col1000);
data ms.ciwc&m;
merge transc1(drop=_name_ col1-col1000) transciw1(drop=_name_ col1-col1000);
proc print data=ms.alldiffd&m;
proc print data=ms.ciwc&m;
proc print data=ms.alldd&m;


run;

ods rtf close;
ods listing;
```

Appendix C - SAS code for Nonparametric Method

```
ods rtf file="d:/nonparametric.rtf";
ods listing close;
libname ms "t:/";
%macro msreportv(seed, repeat, n, m);

%do j=1 %to &repeat;
*data cau_&j(drop=i);
data cau_&j;
%Do i =1 %to &n;
w=rancau(&seed+&j);
output;

%end;
run;

proc sort data=cau_&j;
by w;
run;


data caunew_&j;
set cau_&j;
r=int((&n-1)/2-1.96*sqrt(&n/4));
if _n_=r  then output;
if  _n_=&n-r+1 then output;
data caunew1_&j;
set caunew_&j(drop=r);

proc transpose data=caunew1_&j out=trans_&j;
run;
data trans1_&j(drop=_name_ col1 col2);
set trans_&j;
CIW&m=col2-col1;
if col1<=0 and col2>=0 then c&m=1;
else c&m=0;
run;


%end;

%mend
;

%msreportv(789569,1000,31,68)

proc print data=caunew1_1;

%macro namesv(prefix,maxnum);
%do i=1 %to &maxnum;
```

```
&prefix&i
%end;
;

%mend namesv;

%let m=68;

data ms.alldiffd&m;
set %namesv(trans1_,1000);
run;

data alldd1;
set ms.alldiffd&m(drop=CIW&m);
proc transpose data=alldd1 out=transc;
data transc1;
set transc;
CR=sum(of col1-col1000)/1000*100;
run;
data alldd2;
set ms.alldiffd&m(drop=c&m);
proc transpose data=alldd2 out=transciw;
data transciw2;
set transciw;
aveCIW =mean(of col1-col1000);
medCIW=median(of col1-col1000);
data ms.ciwc&m;
merge transc1(drop=_name_ col1-col1000) transciw2(drop=_name_ col1-col1000);
proc print data=ms.alldiffd&m;
proc print data=ms.ciwc&m;
run;
ods rtf close;
ods listing;
```

Appendix D - SAS code for CR and mean length analysis and plots

```
ods rtf file="t:/cauopCR32.rtf";
ods listing close;
libname ms "t:/";

/*%macro cauchy(m,method, n, p);
data ms.cauaveCIW&m;
set ms.ciwc&m;;
method="&method";
n=&n;
trimp=&p;
run;

%mend;
%cauchy(67,simu,101,0.5)


data ms.cauallaveCIW;
set ms.cauaveCIW1-ms.cauaveCIW8 ms.cauaveCIW20-ms.cauaveCIW27 ms.cauaveCIW40-
ms.cauaveCIW47 ms.cauaveCIW60-ms.cauaveCIW67;
run;

data ms.cauallaveciw;
set ms.cauallaveciw;
aveCIW2=log(aveCIW);
run;*/
proc glm data=ms.cauallaveciw;
class method;
model CR=n method trimp  n*method n*trimp trimp*method n*method*trimp
trimp*trimp/ss3 solution;
output out=resi r=resi p=pred;
run;

proc gplot data=resi;
plot resi*pred;

proc univariate data=resi normal plot;
var resi;
qqplot resi;
run;

symbol1  i=join c = black v=dot;
symbol2  i=join c = red v=sqaure;
symbol3  i=join c = blue v=star;
symbol4  i=join c = purple v=circle;
proc sort data=ms.cauallaveCIW;
```

```
by trimp;
proc gplot data=ms.cauallaveCIW;
plot aveCIW*n=method;
by trimp;
*title 'two way interaction of n*mthod';
run;
proc sort data=ms.cauallaveCIW;
by method;
proc gplot data=ms.cauallaveCIW;
plot aveCIW*trimp=n;
by method;
*title 'two way interaction of n*trimp';
run;
proc sort data=ms.cauallaveCIW;
by n;
proc gplot data=ms.cauallaveCIW;
plot aveCIW*trimp=method;
by n;
*title 'two way interaction of method*trimp';
run;
symbol1  i=join c = black v=dot;
symbol2  i=join c = red v=sqaure;
symbol3  i=join c = blue v=star;
symbol4  i=join c = purple v=circle;
proc sort data=ms.cauallaveCIW;
by trimp;
proc gplot data=ms.cauallaveCIW;
plot CR*n=method;
by trimp;
*title 'two way interaction of n*mthod';
run;
proc sort data=ms.cauallaveCIW;
by method;
proc gplot data=ms.cauallaveCIW;
plot CR*trimp=n;
by method;
*title 'two way interaction of n*trimp';
run;
proc sort data=ms.cauallaveCIW;
by n;
proc gplot data=ms.cauallaveCIW;
plot CR*trimp=method;
by n;
*title 'two way interaction of method*trimp';
run;

ods rtf close;
ods listing;
```

```
******SAS code Chapter 5 Robustness Study
libname ms "d:/warsaw898";
data ms.chap5robust;
set ms.cauallaveciw ms.norallaveciw ms.logisticallaveciw;
run;
data ms.chap5robust2;
set ms.chap5robust;
if method='asym' then delete;
run;
proc sgpanel data=ms.chap5robust2;
panelby n/spacing=5;
vbar dist/response=CR group=trimp;
run;
proc sgpanel data=ms.chap5robust2;
panelby n/spacing=5;
vbar dist/response=aveCIW group=trimp;
run;
```