USE OF THE BOOTSTRAP IN AN ADAPTIVE STATISTICAL PROCEDURE

BY

CHRISTIAN LÉGER

TECHNICAL REPORT NO. 296
JUNE 1988

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS86-00235

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
USE OF THE BOOTSTRAP IN AN ADAPTIVE STATISTICAL PROCEDURE

BY

CHRISTIAN LÉGER

TECHNICAL REPORT NO. 296
JUNE 1988

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS86-00235

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
USE OF THE BOOTSTRAP IN AN ADAPTIVE STATISTICAL PROCEDURE

Christian Léger, Ph.D.
Stanford University, 1988

Numerous estimators in different problems, ranging from the estimation of a finite
dimensional location parameter to the infinite dimensional estimation of a density, are
only specified up to a constant which will be called the smoothing parameter. In most
cases, the optimal choice of that parameter is a function of the underlying distribution of
the observations and of the sample size.

In this dissertation, we will show how the bootstrap can be used to adaptively select
the trimming proportion in trimmed means. Two bootstrap adaptive trimmed means are
introduced. The random trimming proportion is chosen to minimize bootstrap estimates
of the variance, in one case, or the interquartile range, in the other, of the finite sample
distribution of the trimmed means.

They are both shown to be asymptotically normal with an asymptotic variance equal
to the smallest asymptotic variance among the class of trimmed means. Bootstrap con-
fidence intervals for the location parameter based on those bootstrap adaptive trimmed
means, which therefore involve a double bootstrap, are shown to asymptotically have the
claimed coverage probability.

Simulation results demonstrate that the bootstrap adaptive trimmed means have
very good finite sample properties, even for samples of size as small as 10. Finally, other
problems where comparable results are likely to obtain are discussed.
# Table of Contents

1 Introduction .................................................. 1

2 Robustness and Trimmed Means ............................... 7

2.1 Historical Remarks ......................................... 7
2.2 Trimmed Means .............................................. 13
2.3 Notation .................................................. 16

3 Adaptive Trimmed Means .................................... 19

3.1 Jaeckel's Adaptive Trimmed Mean .......................... 19
3.2 Confidence Intervals Based on Jaeckel's Estimator ....... 24
3.3 Bootstrap Adaptive Trimmed Mean ......................... 31
3.4 Confidence Intervals Based on a Bootstrap Adaptive Trimmed Mean 37
3.5 Estimators Based on Alternative Estimators of F ........ 40
3.6 Asymmetric Distributions ................................... 43

4 Proofs ....................................................... 46

4.1 Ideas of The Proofs ......................................... 46
4.2 Proofs .................................................. 53

5 Small Sample Simulations .................................. 88

5.1 Finite Sample Variances ................................... 88
5.2 Bootstrap Confidence Intervals ............................ 104
5.3 Remarks .................................................. 108

6 Possible Extensions ........................................ 118

6.1 Bootstrap Adaptive M-estimation ......................... 118
6.2 Bootstrap Adaptive Trimmed Mean Regression ............ 119
6.3 Bootstrap Adaptive Kernel Density Estimation .......... 121

7 Conclusion .................................................. 123

Bibliography .................................................. 126
Chapter 1

Introduction

This dissertation is about the use of the bootstrap in adaptive statistical procedures. The goal is two-fold: we want to adaptively choose a procedure which is asymptotically optimal in some sense and has good finite sample properties. Before going any further, let's review some of the different approaches to adaptive estimation in the specific setting of the location problem.

Adaptive estimators have been proposed by many authors with different goals in mind. One such goal is asymptotic efficiency. Let's introduce the following notation. Let $X = (X_1, X_2, \ldots, X_n)$ be an i.i.d sample of size $n$ from the distribution $F(x - \theta)$. In this case, one is interested in obtaining an estimator which, without knowing the distribution $F$ that generated the data $X$, will be asymptotically normal with the smallest possible asymptotic variance for that distribution, i.e., Fisher's information lower bound. Stein (1956) first introduced the idea that this goal should be achievable. He gave a simple necessary condition. Many other people worked on the problem in the following years until definitive answers were found by Beran (1974) and Stone (1975). Beran used an adaptive rank estimator whereas Stone introduced the so-called adaptive maximum likelihood estimator. As its name implies, this procedure estimates the density and uses the maximum likelihood estimator of location for that density. Bickel (1982) devoted his Wald lectures to adaptive estimation as is understood here and contains many other references. One problem with such estimators is that finite sample properties are not usually very good until the sample size is large enough, i.e., at least 40.

Another goal in adaptive estimation is finite sample robustness. In this case, one is interested in an adaptive estimator that will have very good finite sample properties
over a fairly broad family of distributions $F$. These estimators are usually constructed as follows. We assume that the distribution of the data is one of $k$ (a small number) distributions $D_i$. For each of them, choose an estimator $T_i$ which has good properties for that distribution. Then construct a selection statistic $s(X)$ whose job is to guess which of the possible distributions generated the data. Divide the range of the selection statistic into $k$ disjoint sets $I_i$, and use the estimator corresponding to the distribution selected by the value of $s(X)$, i.e., if $s(X) \in I_j$ then use $T_j$ as the estimator. Hogg (1967) proposed such an approach to adaptive estimation. In that paper, he used the sample kurtosis as his selection statistic. He then divided the range of the sample kurtosis in $k$ disjoint intervals in such a way that each included the kurtosis of one of the $k$ distributions as an interior point. Hogg (1967) showed that, since the sample kurtosis converges in probability to the kurtosis of the distribution as the sample size increases to infinity, then if the distribution of the data is $D_i$, the asymptotic distribution of this adaptive estimator is the same as that of $T_i$. Therefore if $T_i$ is the maximum likelihood estimator of location of $D_i$, then this adaptive estimator is asymptotically efficient for the $k$ distributions $D_i$'s. Also, it is obvious that if the data is generated by any distribution whose kurtosis is an interior point of $I_i$, again the asymptotic distribution of this adaptive estimator will be the same as that of $T_i$. This is an interesting asymptotic result, but it falls way short of the adaptive estimators of the previous paragraph which are asymptotically efficient for all distributions that satisfy some regularity conditions.

Of course, one could use a different selection statistic. Using a Bayesian approach, Hogg et al. (1972) came up with a selection statistic that basically amounts to pairwise likelihood ratio tests between the different possible distributions. Recently, Cohen and Sackrowitz (1987) showed that a certain modification of that adaptive statistic was generalized Bayes for a certain loss function when using Lebesgue measure on the location parameter $\theta$ as the generalized prior. Needless to say that the loss function assumes that the data was generated by one of the distributions $D_i$'s.

In practice, it is not necessary to explicitly specify the distributions that one has in
mind when devising such an estimator; it suffices to specify a selection statistic and estimators $T_i$'s that should do a good job over a wide range of distributions. Therefore, it is possible to use familiar statistics—such as the mean, the median, and the midmean (25% trimmed mean)—that practitioners feel comfortable with. This is obviously a practical advantage over the methods mentioned previously. Hogg (1974), in a review of adaptive estimation, presented several such adaptive estimators. Another advantage of such methods is their good finite sample properties, even for samples of size 20. By restricting the competing estimators to a small number, the variance incurred by the adaptive step is reduced, while the robustness is guaranteed by the diversity of the estimators. So their properties in small samples along with their simplicity and their intuitive appeal make them attractive to applied practitioners.

On the other hand, one disadvantage of such procedures is the difficulty of objectively assessing their finite sample properties. As we have seen, one way to think about those adaptive estimators is to select certain estimators $T_i$'s that work well for certain distributions $D_i$'s, and then choose a selection statistic that selects which of those distributions best reflects the data at hand. It is therefore natural to assess the behavior of the resulting estimator by looking at its behavior at each of the $D_i$'s. But is it objective? Hogg (1967) introduced, as an example of this type of estimation procedure, an adaptive estimator that used the kurtosis to select from the mean, the 25 percent trimmed mean, the median, and the mean of the observations left out in the computation of the 25 percent trimmed mean. The first three are good for distributions that go from the normal to heavy-tailed distributions, whereas the last one is good for short-tailed distributions. In evaluating that estimator, he looked at its performance for distributions ranging from the uniform to heavy-tailed distributions. Then the Princeton Robustness Study (Andrews et al. (1972)) compared that estimator to many others and found that it was only fair. Hogg (1974) contends that this is due to the fact that they did not consider distributions with tails lower than that of the normal distribution; consequently, the mean of the observations left out in the 25 percent trimmed mean should not even have been included as
one of the possible contenders in the adaptive selection. This clearly illustrates the fact that an objective evaluation of such procedures can be difficult since different objectives suggest different definitions of the adaptive estimator.

The next approach to adaptive estimation can be viewed as a compromise between asymptotic efficiency and finite sample robustness. Jaeckel (1971) used a selection statistic to choose from a family of estimators in such a way that, asymptotically, the adaptive estimator has the same behavior as the best estimator, among those in the family, for the particular distribution from which the data is obtained. In particular, Jaeckel used the family of \( \alpha \)-trimmed means, \( T_n(\alpha; X) \), \( \alpha \in (0, 1/2) \). Since the asymptotic variance of \( T_n(\alpha; X) \) depends both on the value of the trimming proportion \( \alpha \) and the distribution \( F \), it seems only reasonable to use the data to select the trimming proportion, and hence the estimator. The goal is to obtain an adaptive estimator that has an asymptotic variance equal to that of the trimmed mean with the smallest asymptotic variance for the particular distribution \( F \). Clearly this result is weaker than those obtained by Beran (1974) and Stone (1975), but stronger than that of Hogg (1967).

Jaeckel's procedure is as follows. For each \( \alpha \) such that \( n\alpha \) is an integer, compute an estimate of the asymptotic variance of \( T_n(\alpha; X) \), and use the trimming proportion \( \alpha \) that minimizes those estimates. That procedure successfully achieved the stated goal. Furthermore, in small sample simulations, Jaeckel (1971) and Andrews et. al. (1972) showed that Jaeckel's adaptive estimator has very good small sample properties for samples of size as small as 20. One deficiency of the method is that it is not necessarily easy to generalize to other families of estimators because it might be next to impossible to estimate the asymptotic variance. But even for those families of estimators for which one can estimate the asymptotic variance, Jaeckel's approach is not what one would necessarily want to use. Instead of estimating a characteristic of the asymptotic distribution of \( T_n(\alpha; X) \), one should estimate a characteristic of its finite sample distribution, and as was previously done, use the trimming proportion \( \alpha \) that minimizes those estimates. Jaeckel most certainly knew that, but he didn't have the tools to accomplish that task.
But in 1979, Efron introduced a computer intensive method called the bootstrap which allows one to estimate quantities about the finite sample distribution of a host of statistics which, to that date, was an impossible analytical task. The purpose of this dissertation is to show how one can use the bootstrap to construct an adaptive estimator in the same spirit as Jaeckel’s estimator.

Let’s recall the notation. We let $X = (X_1, X_2, ..., X_n)$ be an i.i.d. sample of size $n$ from the distribution $F(x - \theta)$. We are interested in estimating the location parameter $\theta$. We have to choose an estimator among the family of estimators $T_n(\alpha, X)$, $\alpha \in \mathbb{A}$, i.e., we have to choose a value of $\alpha$. To do that, we consider a characteristic $s_n(\alpha, F)$ of the distribution of $T_n(\alpha, X)$. It may either be a characteristic of the finite sample distribution, such as its variance or its interquartile range (hence the subscript $n$), or a characteristic of the asymptotic distribution, such as its asymptotic variance (in which case the subscript could be omitted). We then estimate $s_n(\alpha, F)$ by $s_n(\alpha, \hat{F}_n)$, its bootstrap estimate. Here $\hat{F}_n$ is the empirical distribution function based on the data $X$. Finally, we choose the value $\hat{\alpha}_n$ which minimizes the estimates $s_n(\alpha, \hat{F}_n)$. Our location estimate is then $T_n(\hat{\alpha}_n, X)$.

In a similar spirit, Bell (1980) looked at M-estimates whose scale parameter is chosen so as to minimize the estimated asymptotic variance.

In this dissertation, we concentrate mainly on the $\alpha$-trimmed means. In chapter 2, a brief history of robustness up to Jaeckel (1971) will be presented to better understand the context in which Jaeckel’s estimator was introduced. Then, the trimmed mean and its properties will be introduced. The chapter will conclude with the necessary notation and setup for the other chapters.

Three adaptive trimmed means will be presented in chapter 3. They are Jaeckel’s estimator, and two bootstrap alternatives which minimize estimates of the variance and interquartile range of the finite sample distribution of the $\alpha$-trimmed means. It will be shown that even Jaeckel’s estimator is a bootstrap adaptive trimmed mean. The asymptotic distribution of the three adaptive trimmed means will turn out to be identical and will be shown to minimize the asymptotic variance among the class of $\alpha$-trimmed
means. Finally, the coverage probability of bootstrap confidence intervals based on these three adaptive estimators will be shown to converge to the claimed level of confidence. Note that such a confidence interval involves a double bootstrap, and hence, is computationally very intensive.

Chapter 4 will be devoted to the proofs. An approach unifying the results of the previous chapter will be presented. This will give some hindsight as to what is required, in general, for bootstrap adaptive procedures to succeed.

In chapter 5, the results of a Monte Carlo simulation of the small sample behavior of those three adaptive trimmed means, along with some similar competitors such as jackknife and cross-validation adaptive trimmed means, will be presented. The coverage probabilities of different confidence intervals will also be investigated. Those simulations will demonstrate that bootstrap adaptive trimmed means are successful even in samples of size as small as 10!

It is important to note that even though we have concentrated on the problem of selecting the trimming proportion in a trimmed mean, the idea of the bootstrap adaptive procedure was in no way dependent on that family of estimators, or for that matter, on the location problem. What we need is a family of estimators \( T_n(\alpha, X) \) of a parameter \( \theta \) which is smooth enough in \( \alpha \) and \( F \), where \( F \) is the distribution of the data \( X \). So, in chapter 6 we will talk in general terms about the possible extensions to more interesting and important problems, such as regression and density estimation. Finally, chapter 7 contains our conclusions.
Chapter 2

Robustness and Trimmed Means

In the first chapter, we mentioned some of the different goals that adaptive procedures try to accomplish. We saw that Jaeckel's method is a compromise between asymptotic efficiency and finite sample robustness. So after having spent some time talking about the adaptive aspect of that estimator, this chapter will begin with a brief history of robustness up to Jaeckel (1971), so as to understand how he came up with that estimator. Then, the \( \alpha \)-trimmed means will be presented and their properties reviewed. Finally, we will present the setup and the notation that will be used in the next chapters in describing Jaeckel's adaptive trimmed mean and two alternative adaptive trimmed means based on the bootstrap.

2.1. Historical Remarks

In this section, we shall give a brief history of robustness up to 1971, year of the publication of Jaeckel's adaptive method. The interested reader is referred to Huber (1972), and to chapter 1 of Hampel, Ronchetti, Rousseeuw, and Stahel (1986).

It has long been recognized that, most of the time, samples in the real world do not satisfy the assumption of normality of the observations. Some scientists in the nineteenth century had already noticed that samples often had tails that were longer than those expected from the Gaussian distribution. Some of them even came up with methods that tried to deal with the problem. For instance, the first account of the use of a trimmed mean is found in an article in 1821 by an anonymous author. The author tells us that in some provinces of France, the mean yield of a property of land was estimated from the yields of the previous 20 years by discarding the smallest and largest yields and taking
the mean of the 18 remaining observations.

The problem of dealing with outliers, that is data points far away from the bulk of the data, also has a long history. Scientists, and statisticians in particular, have long been aware that outliers will have a very large influence on statistics such as the mean and the variance. In order to limit their influence, outliers were, at first, subjectively deleted from the data set. Later, "objective" methods were invented to decide whether or not to delete a suspected outlier. Having removed the outliers, the data set was then analysed with the usual estimators, such as the mean. The idea was (and still is for many practitioners) that the "cleaned" data set follows approximately a normal distribution which justifies using the mean.

Then came R. A. Fisher, the founder of mathematical statistics. He has had a major impact on the field of statistics and is the inventor of numerous statistical techniques taught and used everyday by statisticians. For instance, analysis of variance is one of his greatest legacy in the world of applied statistics. The key for the success of techniques such as the analysis of variance is that he was able to obtain the exact distribution of the statistics involved for any finite sample size whenever the data follows the normal distribution. This is very important in the testing of hypotheses about the different effects. The assumption of normality was usually reasonable in the biological data sets that he was encountering, in part as a consequence of the central limit theorem.

So, the mathematical tractability of the normal model, especially for finite sample sizes, along with the (false) sense of security given by the central limit theorem, justified the overall emphasis on the normal model in statistics.

But soon thereafter, statisticians, E. S. Pearson among others, wondered about what happens to the distribution of the test statistics when the assumption of normality is violated. They were mainly interested in determining whether the different tests are conservative, what is now known as robustness of validity. In fact, it was in the study of such tests that G. E. P. Box coined the term "robustness" in Box (1953). They discovered that some are not, such as the test of equality of variances, while others are, such as the
t-test. But the more difficult problem of robustness of efficiency, that is the behavior of the power function, was rather neglected at the time.

The estimation problem did not get the same kind of attention. We mentioned earlier the problem of gross errors. But what about small deviations from normality? It was either hoped that small deviations would have small effects so that it would be reasonable to use the optimal estimators for the normal distribution, or the problem was ignored. A first indication that the mean and variance might not be efficient when the distribution differs slightly from a normal came when Fisher (1922) stressed their inefficiencies in all but a very small region of Pearson’s curves around the normal. But his argument was directed against K. Pearson’s method of moments. Obviously, the same argument is just as valid against the uncritical use of least squares methods when the normality assumption is in doubt. In any case, dangers related to point estimation in slightly nonnormal settings remained fairly unnoticed for decades.

In fact, it was only in 1960 that such dangers became fully appreciated in the landmark article of Tukey (1960). To be more precise, a large part of that work was done in the fifties by the Statistical Research Group at Princeton and ended up in a number of memorandums that were not widely available. In that article, Tukey studies the properties of various location and scale estimators when the random variables are distributed according to a contaminated normal distribution, that is, each observation is either a standard normal or a normal with a standard deviation $h$, with probabilities $1 - \gamma$ and $\gamma$, respectively. For instance, he looked at the classic problem of whether to use the mean absolute error or the square root of the mean squared error as an estimate of the scale of a distribution. Eddington was a proponent of the mean absolute error. But Fisher (1920) showed that for the normal distribution, the square root of the mean squared error was a more efficient estimator of scale than the mean absolute error. As a matter of fact, in this case the asymptotic relative efficiency of the mean absolute error to the square root of the mean squared error is only 0.876. Nevertheless, Eddington maintained that the mean absolute error was a more precise estimator of scale based on experience with astronom-
ical data. Tukey gave considerable weight to that belief by noticing that the asymptotic relative efficiency is 1.0 when the observations come from a contaminated normal with $\gamma$ just less than 0.008 and $h$, the standard deviation for the wider normal, equal to 3. This means that out of a sample of size 1000, 8 observations are expected to be from the wider constituent. But since only 40% of the observations from that distribution should be outside of the interval $(-2.5, 2.5)$, only 2 observations, on average, from the wider normal will be on each side of those two tails. That means that 2 outliers, on average, in each tail in a sample of size 1000 is sufficient to eliminate the 12% advantage of the mean squared error. He then proceeded to show that if $\gamma$ was higher than 0.008, the mean absolute error is more efficient! It is interesting to note that Fisher was the only statistician queried by Tukey who expected such a drastic effect for such a small value of $\gamma$.

In hypothesis testing, robustness of validity guarantees that the significance level is at least as large as what is claimed. There is of course no equivalent concept in point estimation. But with that work, Tukey put the emphasis on robustness of efficiency, that is, estimators should be relatively efficient over small departures from the normality assumption. He also laid down some of the foundations of the field of robustness as it is known today. Many of the (somewhat) arbitrary choices that he made in this study have been kept to this day by a host of investigators in robustness. For instance, he mentioned that as a first step only, one should study the location problem for symmetric distributions; with the symmetry assumption, the location of the distribution is well defined and unambiguous. This explains, in part, why so much of the robustness literature has assumed symmetry of the underlying distribution. Also computers were not as prevalent in those days. Yet Tukey suggested that much could be learned about the small sample properties of different estimators by "experimental sampling", better known today as a Monte Carlo experiment. Well, Tukey's words didn't fall on deaf ears. In fact, Stigler (1977) noted that: "As the costs of computation have declined in recent years, the use of the computer for the production of pseudo-random numbers has increased, to the point where the volume of synthetic data produced in universities and research laboratories may even surpass
the Census Bureau's production of the real thing."

As the title of the article suggests, Tukey was working with contaminated distributions. He did not introduce the idea; it had already been studied in numerous publications. But from his experience with largish samples, he favored the use of normals with the same mean and different variances. Specifically, he chose to use a standard deviation three times as large for the second component \( h = 3 \). That choice has been followed by a host of studies thereafter. In fact, it is the ancestor of the one-wild sampling scheme which consists of sampling one observation from a normal with standard deviation equal to three while the remaining observations are sampled from the standard normal distribution.

Through theoretical considerations and his own experience, Tukey was lead to consider alternatives to the Gaussian distribution that have longer tails. This, once again, explains in part why so many robustness Monte Carlo studies—including the most famous of all, the Princeton Robustness Study (Andrews et. al. (1972))—only considered distributions with tails at least as long as the normal.

Another of Tukey's legacies is the revival of the \( \alpha \)-trimmed means, the old French custom previously mentioned. It is interesting to note how they were introduced in the paper. Assuming that the amount of contamination, \( \gamma \), and the ratio of the standard deviations of the two normals, \( h \), is known, Tukey decided to approximate the maximum likelihood estimator by an L-estimator. Then he noticed that this L-estimator could itself be well approximated by an appropriate \( \alpha \)-trimmed mean. Having introduced the trimmed means, he then went on to compute their asymptotic efficiencies for the contaminated normals. He noticed that a small amount of trimming leads to a small reduction of efficiency at the normal distribution, while achieving big gains even for a small amount of contamination.

Tukey and McLaughlin (1963) pursued the program set forth in Tukey (1960) by suggesting an alternative to the t-test based on \( \alpha \)-trimmed means. This paper is not of great importance in the history of robustness per se, but it is a significant link to Jaeckel (1971) as we will soon discover. They began by noting that the t-test was made of two
pieces. The numerator is an estimator of location while the denominator is an estimator of the variability of the numerator. They also noticed that, in the t-test, the ratio of the mean of the square of the denominator to the variance of the numerator is equal to 1 for all distributions. In devising a more robust test, they decided to use a more robust estimate of location in the numerator (a trimmed mean) and then to guess a denominator which would satisfy the two requirements just mentioned, namely, that it should be an estimator of the variability of the trimmed mean, and that, over a reasonably broad spectrum of distributions, the mean of the squared denominator and the variance of the numerator should be in constant ratio. This would of course be verified through Monte Carlo simulations. After some intuitive considerations and different experiments, they came up with a denominator which, it turns out, is almost the square root of a "plug-in" estimate of the asymptotic variance of the α-trimmed mean. It is interesting to note that the first published statement of the asymptotic normality of the trimmed mean is in Bickel (1965). But Bickel mentions that it was stated in an unpublished technical report of Harris and Tukey (1949). Yet, Tukey and McLaughlin make no mention of this result in their search for a denominator.

In that paper, they also suggested what is at the heart of this dissertation: the use of the data to choose the trimming proportion α. They proposed choosing α < A(n) which minimizes the value of the denominator associated with the α-trimmed mean, where n is the sample size and A(n) is a function that grows with n. They noted that the estimates of the variance of the α-trimmed mean, being based on a Winsorized sum of squares, would be too unstable for values of α too close to 1/2. Therefore they decided to exclude them from consideration by restricting the possible values of α to be smaller than A(n). This is exactly the program that Jaekel tackled as will be seen in the next chapter.

One of the most influential contributions to the theory of robustness, as it is known today, is that of Huber (1964). In that paper, Huber presented his minimax approach to robustness. The idea consists of choosing an estimator which minimizes the maximum asymptotic variance as the distribution of the observations ranges in a (symmet-
Section 2.2: Trimmed Means

ric) neighborhood of the normal distribution. Huber gave an explicit solution to that problem. Moreover, in the process of finding the solution, he introduced a whole class of estimators—the M-estimators. This is a rich class of estimators that includes the mean, the median and the maximum likelihood estimators.

Of importance here, is the fact that the explicit solution to that minimax problem contains a tuning parameter which depends on the assumed size of the neighborhood of distribution functions. In his "Proposal 3", Huber suggested to select the parameter that minimizes the estimated asymptotic variance. This is in the same spirit as Tukey and McLaughlin, and this is again the idea that will be used by Jaeckel.

In these brief historical remarks, I have clearly left out some important contributions, for instance that of Hampel (1971). But the goal was to give an overview of the ideas leading to Jaeckel's contribution. Along with these remarks about robustness, one must not forget the discussion of adaptive inference contained in the previous chapter. In fact, it is difficult to classify Jaeckel's estimator in either the theory of robustness or the theory of adaptive inference. For the same reason, this dissertation will sometimes concentrate on ideas of robustness, while other times concentrate on adaptiveness. Now let's review some of the properties of the $\alpha$-trimmed mean.

2.2. Trimmed Means

In this section, the $\alpha$-trimmed mean will formally be introduced. Two asymptotically equivalent definitions will be given and we will state the most general result on its asymptotic behavior, a theorem due to Stigler (1973).

Let $X_1, X_2, \ldots, X_n$ be identically and independently distributed according to the distribution function $F$. Let $\hat{F}_n$ be the empirical distribution function of the sample, i.e.,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leq x\}, \quad -\infty < x < \infty,$$

where $I$ is the indicator function.
**Definition 2.1:** The (symmetric) $\alpha$-trimmed mean $T_\alpha^n$ is given by:

$$T_\alpha^n = (1 - 2[n\alpha])^{-1} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)},$$

where $\alpha \in [0, 1/2)$, $[\cdot]$ represents the greatest integer function, and $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ are the order statistics.

Note that $T_\alpha^n$ is the mean and that if $\alpha \in \left([\frac{n+1}{2}] - 1/n, 1/2\right)$, then $T_\alpha^n$ is the median. It is important to note that $T_\alpha^n$ takes only $([\frac{n-1}{2}] + 1)$ different values as $\alpha$ ranges from 0 to 1/2. Let's introduce the following notation before we present the asymptotic distribution of $T_\alpha^n$.

Let $a^\alpha$ and $b^\alpha$ be the largest $\alpha^{th}$ and smallest $(1 - \alpha)^{th}$ percentiles of $F$; that is,

$$a^\alpha = \sup\{ x : F(x) \leq \alpha \},$$
$$b^\alpha = \inf\{ x : F(x) \geq 1 - \alpha \}.$$

Let $A^\alpha$ and $B^\alpha$ be the lengths (possibly zero) of the intervals of the $\alpha^{th}$ and $(1 - \alpha)^{th}$ percentiles:

$$A^\alpha = a^\alpha - \inf\{ x : F(x) \geq \alpha \},$$
$$B^\alpha = \sup\{ x : F(x) \leq 1 - \alpha \} - b^\alpha.$$

Further, define

$$G^\alpha(x) = \begin{cases} 
0, & \text{if } x < a; \\
(F(x) - \alpha)/(1 - 2\alpha), & \text{if } a \leq x < b; \\
1, & \text{if } x \geq b.
\end{cases}$$

and set

$$\mu^\alpha = \int_{-\infty}^{\infty} x dG^\alpha(x),$$
$$\sigma^2(\alpha) = \int_{-\infty}^{\infty} x^2 dG^\alpha(x) - (\mu^\alpha)^2.$$

**Theorem 2.1:** (Stigler (1973)). The distribution of $n^{1/2}(T_\alpha^n - \mu^\alpha)$ converges weakly to the distribution of $Z^\alpha$, where $Z^\alpha = (1 - 2\alpha)^{-1}[Y_1^\alpha + (b^\alpha - \mu^\alpha)Y_2^\alpha + (a^\alpha - \mu^\alpha)Y_3^\alpha + B^\alpha \max(0, Y_2^\alpha) - A^\alpha \max(0, Y_3^\alpha)]$ and $E(Z^\alpha) = [(B^\alpha - A^\alpha)(\alpha(1-\alpha))^{1/2}] / [(1-2\alpha)(2\pi)^{1/2}],$. 


where $Y_1^\alpha$ is $N(0, (1-2\alpha)\sigma^2(\alpha))$ and independent of $(Y_2^\alpha, Y_3^\alpha)$, and $(Y_2^\alpha, Y_3^\alpha)$ is $N(0, C^\alpha),

\[ C^\alpha = \begin{pmatrix} \alpha(1-\alpha) & -\alpha^2 \\
-\alpha^2 & \alpha(1-\alpha) \end{pmatrix}. \]

As an important special case, suppose that $F$ is symmetric about $\theta$, that is, $F(x-\theta) + F(\theta-x) = 1$, and that $F^{-1}$ is continuous at $\alpha$ and $1-\alpha$, where $F^{-1}(x) = \inf\{y : F(y) \geq x\}$. Then $A^\alpha$ and $B^\alpha$ are equal to zero and we get the following corollary.

**Corollary 2.1:** For $\alpha \in [0, 1/2)$, let $F$ be symmetric about $\theta$ and let $F^{-1}$ be continuous at $\alpha$ and $1-\alpha$. Then as $n \to \infty$, the distribution of $n^{1/2}(T_n^\alpha - \theta)$ converges weakly to the distribution of $Z^\alpha$, where $Z^\alpha$ is $N(0, \sigma^2(\alpha, F))$ and

\[ \sigma^2(\alpha, F) = (1-2\alpha)^{-2} \left[ \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} (x-\theta)^2 dF(x) + 2\alpha(F^{-1}(\alpha) - \theta)^2 \right]. \tag{2.2} \]

It is important to note that the theorem is valid for all distribution functions. It is also interesting to note that the trimmed mean will not be asymptotically normal if either the $\alpha^{th}$ or $(1-\alpha)^{th}$ quantile is not unique. Therefore the asymptotic distribution of the $\alpha$-trimmed mean depends strongly on some local characteristics of the distribution $F$.

Now, let us introduce an alternative definition of the trimmed mean based on a functional.

**Definition 2.2:** For any $\alpha \in [0, 1/2)$, consider the functional

\[ T(\alpha, F) = (1-2\alpha)^{-1} \int_\alpha^{1-\alpha} F^{-1}(t)dt. \]

If $\hat{F}_n$ is the empirical distribution function of a sample, then $T(\alpha, \hat{F}_n)$ will also be called the $\alpha$-trimmed mean of the sample.

Although the definitions look very different, they are very similar. They both agree for $\alpha = i/n$ with $i = 0, 1, \ldots, \left[\frac{n-1}{2}\right] - 1$. Asymptotically, the two definitions agree for all $\alpha \in [0, 1/2)$ so that asymptotic results for the first definition can be worked out using
the second definition. The fact that $T(\alpha, \cdot)$ is a functional allows us to use Von Mises calculus which will greatly simplify asymptotic calculations. Consequently, even though we will use the first definition in finite samples, the second one will be used in asymptotic calculations.

Now we are in a position to introduce the notation that will be used throughout the rest of this dissertation.

### 2.3. Notation

Let $X_1, X_2, \ldots, X_n$ be identically and independently distributed according to the distribution function $F$. Let $\hat{F}_n$ be the empirical distribution function given by (2.1). We have already introduced the $\alpha$-trimmed mean $T(\alpha, \hat{F}_n)$. Now let’s introduce the trimmed mean process $M_n(\alpha, \hat{F}_n, F)$ given by

$$M_n(\alpha, \hat{F}_n, F) = n^{1/2}[T(\alpha, \hat{F}_n) - T(\alpha, F)], \quad \alpha \in [0, 1/2). \quad (2.3)$$

Here, $T(\alpha, F)$ defines the location parameter of interest. Note that if the distribution $F$ is symmetric about $\theta$, then $T(\alpha, F) = \theta$ for all $\alpha \in [0, 1/2)$ and so (2.3) can be rewritten as

$$M_n(\alpha, \hat{F}_n, F) = n^{1/2}[T(\alpha, \hat{F}_n) - \theta], \quad \alpha \in [0, 1/2).$$

As will be seen in the next chapter, we will assume that $F$ is symmetric so that the location of the distribution is unambiguously defined.

Let $L_n(x; \alpha, F)$ be the distribution function (evaluated at $x$) of the trimmed mean process (treating $\alpha$ as fixed). Then,

$$L_n(x; \alpha, F) = \text{Prob}_F\{M_n(\alpha, \hat{F}_n, F) \leq x\}. \quad (2.4)$$

Also, let $L(x; \alpha, F) = \Phi(x/\sigma(\alpha, F))$ where $\Phi(\cdot)$ is the distribution function of the standard normal and $\sigma(\alpha, F)$ is given by (2.2). Then corollary 2.1 can be restated as follows: $L_n(x; \alpha, F) \rightarrow L(x; \alpha, F)$, as $n \rightarrow \infty$ for each $\alpha \in [0, 1/2)$. Of course, this result is not a result about the behavior of the trimmed mean process per se, because only one value of
$\alpha$ is considered at a time. But since we will be looking at bootstrap confidence intervals for adaptive trimmed means, we need a more global result that will demonstrate that the trimmed mean process is smooth in $\alpha$ and $F$. This result will be presented in chapter 4. Now let's consider adaptive choices of the trimming proportion $\alpha$.

In general, an adaptive choice is just a random variable $\alpha(X_1, X_2, \ldots, X_n)$ taking values in $[0, 1/2)$. But we will be interested in the following setup. Consider a characteristic $S_n(\alpha, F)$ of the distribution of $T(\alpha, \hat{F}_n)$, usually a measure of spread. Note that by the distribution of $T(\alpha, \hat{F}_n)$, we mean either the finite sample distribution or the asymptotic distribution, in which case the subscript $n$ in the functional $S_n$ could, in principle, be suppressed. The goal is then to estimate $S_n(\alpha, F)$. The bootstrap methodology, introduced by Efron (1979), offers a natural estimate of functionals such as $S_n(\alpha, F)$, namely $S_n(\alpha, \hat{F})$, where $\hat{F}$ is an estimate of $F$. In general, the empirical distribution function $\hat{F} = \hat{F}_n$ is used. Often one has to resort to Monte Carlo simulation to approximate $S_n(\alpha, \hat{F}_n)$. So, coming back to our problem, we then estimate $S_n(\alpha, F)$ by its bootstrap estimate $S_n(\alpha, \hat{F}_n)$, for each $\alpha \in A_n$, where $A_n = \{\alpha : \alpha \in [\alpha_0, 1/2 - b_n] \text{ and } n\alpha \in \mathbb{N}\}$, and $\alpha_0$ and $b_n$ are constants between 0 and 1/2. Note that the lower bound $\alpha_0$ is fixed whereas the upper bound $1/2 - b_n$ is allowed to vary with the sample size. Typically, the sequence $\{b_n\}$ will either be constant, or tend to 0 as $n \to \infty$. Also, recall that according to definition 2.1, the estimator $T(\alpha, \cdot)$ only needs to be evaluated at the $[\frac{n-1}{2}] + 1$ different values of $\alpha$ such that $n\alpha$ is an integer. Therefore, by the nature of the trimmed means, the restriction to $\alpha$'s such that $n\alpha$ is an integer is not a serious one. Having computed bootstrap estimates of a characteristic of the distribution of the $\alpha$-trimmed mean, we take the value of $\alpha$ which minimizes those estimates, namely,

$$\alpha_n(\hat{F}_n) = \arg \min_{\alpha \in A_n} S_n(\alpha, \hat{F}_n).$$

The adaptive trimmed mean estimate of location is then $T_n(\hat{F}_n) = T(\alpha_n(\hat{F}_n), \hat{F}_n)$. Since more than one such adaptive trimmed mean will be introduced, we will use superscripts to differentiate among the different $S_n$, $\alpha_n$, and $T_n$. 
In dealing with bootstrap confidence intervals for the location parameter \( \theta \), it will be necessary to establish the weak convergence of the distribution of \( T_n(\cdot) \) where the distribution of the observations is random and varies with \( n \). Therefore, it will be more convenient to work with the distribution function given by

\[
J_n(x, F) = \text{Prob}_F\{n^{1/2}[T_n(\hat{F}_n) - T(\alpha_n(\hat{F}_n), F)] \leq x\}
= \text{Prob}_F\{n^{1/2}[T(\alpha_n(\hat{F}_n), \hat{F}_n) - T(\alpha_n(\hat{F}_n), F)] \leq x\}. \tag{2.5}
\]

Having introduced all that notation, we are now in a position to turn on to chapter 3 where Jaeckel's adaptive trimmed mean and the other two alternatives will be studied.
Chapter 3

Adaptive Trimmed Means

In this chapter, three adaptive estimators of location based on trimmed means will be considered: one due to Jaeckel (1971), and two alternatives based on the bootstrap. For each estimator asymptotic normality will be established with an asymptotic variance equal to the smallest possible asymptotic variance among the class of trimmed means. Then we will show that it is possible to use the bootstrap to construct confidence intervals for the location parameter that will asymptotically have the claimed coverage probability. Since the alternative estimators are based on the bootstrap and that the above results apply to the usual bootstrap, i.e., when the distribution function $F$ is estimated by the empirical distribution function $\hat{F}_n$, the natural question is then what if another estimator of $F$ is used? Do the results still apply? It will be shown that if one uses a kernel distribution function estimator or a symmetric estimator that satisfies certain conditions, the results remain valid. Finally, the case of asymmetric distribution functions will be discussed in the last section. The proofs will be delayed until next chapter.

3.1. Jaeckel’s Adaptive Trimmed Mean

Let’s consider the setup introduced in section 2.3 of chapter 2. Let

$$S^{\text{Jae}}(\alpha, F) = \sigma^2(\alpha, F),$$

where $\sigma^2(\alpha, F)$ is the asymptotic variance of the $\alpha$-trimmed mean of a distribution $F$ given by (2.2). Jaeckel estimates $S^{\text{Jae}}(\alpha, F)$ by $S^{\text{Jae}}(\alpha, \hat{F}_n)$, where $\hat{F}_n$ is the empirical
distribution function of a sample of size \( n \) from \( F \). Specifically, we have

\[
S^{Jae}(\alpha, \hat{F}_n) = \frac{1}{(1 - 2\alpha)^2} \left\{ \frac{1}{n} \sum_{i=\lfloor an \rfloor + 1}^{n-\lfloor an \rfloor} [X(i) - T(\alpha, \hat{F}_n)]^2 + \alpha[X(\lfloor an \rfloor + 1) - T(\alpha, \hat{F}_n)]^2 + \alpha[X(n-\lfloor an \rfloor) - T(\alpha, \hat{F}_n)]^2 \right\},
\]

(3.1)

where \( \lfloor \cdot \rfloor \) denotes the greatest integer function. If you take a careful look at (2.2), you will notice that (3.1) is indeed a bootstrap estimate of (2.2). Of course this formula does not require any Monte Carlo simulation; it is an analytical bootstrap. Then for

\[
A_n = \{ \alpha : \alpha \in [\alpha_0, 1/2 - b_n], \text{ and } n\alpha \in \mathbb{N} \},
\]

(3.2)

let

\[
\alpha_n^{Jae}(\hat{F}_n) = \arg\min_{\alpha \in A_n} S^{Jae}(\alpha, \hat{F}_n),
\]

(3.3)

and denote \( T(\alpha_n^{Jae}(\hat{F}_n), \hat{F}_n) \), Jaekel’s adaptive trimmed mean estimate of location by \( T_n^{Jae}(\hat{F}_n) \). Finally, let \( J_n^{Jae}(x, F) \) be the centralized distribution function of \( T_n^{Jae}(\hat{F}_n) \), namely,

\[
J_n^{Jae}(x, F) = \text{Prob}_F\{n^{1/2}[T_n^{Jae}(\hat{F}_n) - T(\alpha_n^{Jae}(F), F)] \leq x\}
\]

\[
= \text{Prob}_F\{n^{1/2}[T_n^{Jae}(\hat{F}_n) - \theta] \leq x\},
\]

the second equality being valid whenever \( F \) is symmetric about \( \theta \).

Computationally, this estimator is not very difficult to compute. One starts by sorting the data. Then one computes the trimmed mean with the largest trimming proportion, along with the estimate of its asymptotic variance. The other trimmed means (working down to the smallest trimming proportions) and their corresponding estimates of asymptotic variance are then computed. Even though the trimmed mean can be updated at each iteration, the estimate of the asymptotic variance can’t be. Therefore, the bulk of the time is taken up by calculating the asymptotic variance which is \( O(n^2) \) (since there are \( O(n) \) of them to compute), whereas the sorting step is \( O(n \log n) \), where \( n \) is the sample size.

As was alluded to in the previous chapter, Jaekel introduced the estimator \( T_n^{Jae}(\hat{F}_n) \) as a robust estimator. As we have seen, he was following a recommendation of Tukey and
McLaughlin (1963). Notwithstanding that fact, there are many good reasons to use the trimmed means as the class of estimators to choose from. It is a simple class; a trimmed mean is an intuitive robust procedure that is often used in practice. For instance, the overall mark in sports such as figure skating and diving is a trimmed mean of the scores of all the judges. So even though there might exist better classes of estimators to consider, this is a good class to start with.

Note that (3.1) is a Winsorized estimate of variance, centered at the trimmed mean. By that we mean that it is a mean of squared differences between the order statistics and a trimmed mean, replacing the terms corresponding to the \( n\alpha \) lower and upper order statistics with the term corresponding to \( X_{([n\alpha]+1)} \) and \( X_{(n-[n\alpha])} \), respectively. Winsorized means were also considered by Tukey and McLaughlin. It is interesting to note that the asymptotic variance of a trimmed mean is a Winsorized variance. We mentioned that (3.1) was very similar to what Tukey and McLaughlin had suggested. As a matter of fact, they had suggested using a Winsorized estimate of variance, centered at the Winsorized mean instead of the trimmed mean. Now let’s introduce the assumptions required for the asymptotic normality of this estimator.

**Assumption A:** Consider a distribution function \( F \) and let \( 0 < \alpha_0 < 1/2 \) be fixed.

(A.1) The distribution function \( F \) is symmetric about an unknown parameter \( \theta \), and it has a density \( f \).

(A.2) For some \( 0 < \epsilon_0 < \alpha_0 \), and \( f_0 > 0 \), \( f(x) \geq f_0 \) on \( \{ x : \alpha_0 - \epsilon_0 \leq F(x) \leq 1 - \alpha_0 + \epsilon_0 \} \).

(A.3) \( \sigma^2(\alpha, F) \) has a unique global infimum over the set \( A \) where \( A = [\alpha_0, 1/2] \) if \( b_n \to 0 \) in (3.2) and (3.3), or \( A = [\alpha_0, \alpha_1] \) if \( b_n \) is constant and \( \alpha_1 = 1/2 - b_n \). Then let \( A_0(F) = \arg\inf_{\alpha \in A} \sigma^2(\alpha, F) \).

(A.4) \( f(\theta) \) is finite.

(A.5) \( A_0(F) \neq 1/2 \).

Let’s take a careful look at these assumptions. The assumption of symmetry is a very strong assumption. There are often good extraneous reasons to assume symmetry of
the observations, such as when one is dealing with measurement errors. But many other problems simply are not symmetric. In such a case, the location of the distribution is not well defined, as for instance, the mean and the median are not equal. So, it might be argued that the parameter that an $\alpha$-trimmed mean is estimating is $T(\alpha, F)$, as given by definition 2.2. But then, each member of the class of trimmed means is estimating a different characteristic of the distribution $F$. Thus, an adaptive trimmed mean not only chooses an estimator at random, but it also chooses the parameter that it is estimating at random. Hence, for simplicity of interpretation, we follow the lead of Tukey (1960), as most authors in robustness have done, and we assume symmetry. Fortunately, the actual results do not depend on that assumption at all (except for theorem 3.8) and section 3.6 will discuss the problems of interpretation when the underlying distribution is asymmetric.

The assumption that $F$ has a density $f$ is a necessary evil. It is either more or less restrictive than the assumption of symmetry, depending on the point of view. If one had a large data set, one could test the hypothesis of symmetry to decide whether to go ahead and use this method, or exercise care and possibly use another method. On the other hand, even with a large data set, one cannot test the assumption that $F$ has a density. Hence, one can hardly ever feel, with any degree of certainty, that the data at hand was generated by a distribution which admits a density. But then again, I don't know of many data analysts that worry too much whether or not this assumption is met, whereas quite a few do worry about the symmetry assumption. Recall that this assumption is not necessary for theorem 2.1 and corollary 2.1 to apply. Nevertheless, this assumption, which is tied to (A.2), will make life much easier when our results are proven in the next chapter.

Assumption (A.2) has to do with uniformity of the weak convergence of the trimmed means to normality. We have seen in theorem 2.1 that a necessary and sufficient condition for the asymptotic distribution to be normal is that the inverse distribution function be continuous, i.e., that $f$ not be 0 in an interval about $F^{-1}(\alpha)$ and $F^{-1}(1 - \alpha)$. But assumption (A.2) is stronger in that it requires that $f$ be bounded below by a positive value uniformly on $\{ x : \alpha_0 - \epsilon_0 \leq F(x) \leq 1 - \alpha_0 + \epsilon_0 \}$. That assumption is necessary
to insure that the asymptotic convergence will be uniform which allows us to select the trimming proportion adaptively.

Assumption (A.3) is not very crucial. If the global maximum (over either \([\alpha_0, 1/2]\) or \([\alpha_0, \alpha_1]\)) is attained by more than one point, then a simple modification will give us the same result. For instance, look at remark 4.3 for a simple modification which will guarantee that the method will also be asymptotically normal with the smallest possible asymptotic variance among the family of trimmed means. Note that by assumption (A.2) and (2.2), \(\sigma^2(\alpha, F)\) is continuous in \(\alpha\), and so assumption (A.3) implies that \(A_0(F)\) exists and is well-defined.

The assumption that the density at the median be finite (A.4) will be needed in the next section by the bootstrap adaptive trimmed mean based on the interquartile range.

Assumption (A.5) is a sufficient condition to guarantee the weak convergence of the adaptive trimmed mean. It is not clear what happens when it is not met. For more details, see chapter 4.

Finally, note that it is assumed that \(\alpha_0\), the smallest trimming proportion considered, is strictly positive. That assumption rules out getting the best possible result for one of the most important distributions, namely, the normal. More precisely, it means that, with this proof at least, one can’t get an asymptotic variance equal to the smallest asymptotic variance over \([0, 1/2]\), because that minimum is achieved at 0. But then again, we can get arbitrarily close by choosing \(\alpha_0\) arbitrarily close to 0. Nevertheless, in the next chapter we will see why that assumption seems necessary and how it might possibly be relaxed. See remark 4.2.

Now let’s see the properties of Jaeckel’s estimate.

**Lemma 3.1**: Let \(\Gamma\) satisfy assumptions (A.1) through (A.3). Let \(n^{1/4}b_n \to \infty\) as \(n \to \infty\). Then

\[
\alpha_n^{\text{Jac}}(\hat{F}_n) \to A_0(F) \quad \text{in probability.}
\]

This result is one of the key in proving the following theorem.
THEOREM 3.1: Let $F$ satisfy assumptions (A.1) through (A.3). Let $n^{1/4}b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let's further assume (A.5). Then
\[
\sup_x |J_n^{Jae}(x, F) - \Phi(x\sigma(A_0(F), F))| \to 0
\]
as $n \to \infty$ where $\Phi(x)$ is the distribution function of the standard normal distribution, and $\sigma^2(\alpha, F)$ is given by (2.2).

In other words, Jaekel's estimator is asymptotically normal with an asymptotic variance equal to $\sigma^2(A_0(F), F)$. Note that, by definition, $\sigma^2(\alpha, F)$ achieves its minimum (over $[\alpha_0, 1/2]$ or $[\alpha_0, \alpha_1]$) at $A_0(F)$. Therefore, this adaptive trimmed mean does asymptotically as well as the best trimmed mean (with fixed trimming proportion) for distribution $F$, namely, $T(A_0(F), \hat{\theta}_n)$. Of course this result does not necessarily imply that its (standardized) finite sample variance converges to that asymptotic variance, and therefore, that this estimator will have the smallest possible variance as $n \to \infty$. Moreover, this result says little about its finite sample distribution for small sample sizes. We will come back to small sample results in chapter 5. Note also that under assumptions (A.1) and (A.2) that $\sigma^2(\alpha, F)$ is a continuous function in $\alpha$, (for $\alpha \geq \alpha_0$), so that
\[
\sigma^2(1/2, F) = \lim_{\alpha \to 1/2} \sigma^2(\alpha, F) = 1/\{4f^2(\theta)\}
\]
which is the asymptotic variance of the median. This is important when $A_0(F) = 1/2$ as in the case of the double exponential distribution.

The proof of this result for $b_n$ equal to the same constant for all $n$ is in Jaekel (1971). The more general version contained here can be extracted from carefully bounding each error term in that proof. Nevertheless, the proof of the next theorem will also apply to theorem 3.1. Now, let's consider confidence intervals for the location parameter $\theta$ based on that estimator.

3.2. Confidence Intervals Based on Jaekel's Estimator

If the finite sample distribution of $T_n^{Jae}(\hat{\theta}_n)$ were known, then an exact confidence interval for the location parameter $\theta$ could be constructed. Since it is not the case,
one has to resort to approximations to $J_n^{\text{Jac}}(x, F)$. Of course, one could in principle use randomization tests to obtain an exact confidence interval based on $T_n^{\text{Jac}}(\hat{F}_n)$. This test would consist of all values $\theta_0$ such that a randomization test of the null hypothesis $\theta = \theta_0$ is not rejected at the desired level. But such a confidence interval could not be implemented exactly as one cannot compute an infinite number of randomization tests. Even an approximation would be highly computer intensive, at least as much as the approximate confidence interval based on the bootstrap which will soon be described.

The usual approximation to the finite sample distribution of $T_n^{\text{Jac}}(\hat{F}_n)$ consists of using an appropriate normal distribution, i.e., approximating $J_n^{\text{Jac}}(x, F)$ by $\Phi(x \tau)$ for a certain fixed positive value of $\tau$. Let $Z(\tau)$ be distributed according to a normal distribution with mean 0 and variance $\tau^2$. In order to construct a $1 - 2\beta\%$ two-sided confidence interval for the location parameter $\theta$, consider the following equality:

\[
1 - 2\beta = \text{Prob}_F\{x_{\beta}^{\text{Jac}}(F) \leq n^{1/2}[T_n^{\text{Jac}}(\hat{F}_n) - \theta] \leq x_{1-\beta}^{\text{Jac}}(F)\}; \tag{3.4}
\]

\[
= \text{Prob}\{\tau z_{\beta} \leq Z(\tau) \leq \tau z_{1-\beta}\},
\]

where $z_{\beta} = \Phi^{-1}(\beta)$ is the $\beta^{\text{th}}$ percentile of the standard normal distribution and $x_{\beta}^{\text{Jac}}(F) = J_n^{\text{Jac}}^{-1}(\beta, F)$ is the $\beta^{\text{th}}$ percentile of the distribution of $n^{1/2}[T_n^{\text{Jac}}(\hat{F}_n) - \theta]$. The approximation then consists of inverting the probability statement (3.4) and using $\tau z_{\beta}$ instead of $x_{\beta}^{\text{Jac}}(F)$, and similarly for the upper bound giving the following approximate confidence interval for $\theta$:

\[
\theta \in [T_n^{\text{Jac}}(\hat{F}_n) - n^{-1/2}\tau z_{1-\beta}, T_n^{\text{Jac}}(\hat{F}_n) - n^{-1/2}\tau z_{\beta}]. \tag{3.5}
\]

In order to use that approximation, the approximate variance $\tau^2$ must be chosen. A natural choice would be to use the asymptotic variance $\sigma^2(A_0(F), F)$, but since $F$ is unknown, one has to resort to $\sigma^2(\alpha_n^{\text{Jac}}(\hat{F}_n), \hat{F}_n)$, i.e., $S_n^{\text{Jac}}(\alpha_n^{\text{Jac}}(\hat{F}_n), \hat{F}_n)$. The approximate confidence interval based on the asymptotic normality of Jaeckel's estimator is then given by

\[
\theta \in [T_n^{\text{Jac}}(\hat{F}_n) - n^{-1/2}S_n^{\text{Jac}}(\alpha_n^{\text{Jac}}(\hat{F}_n), \hat{F}_n)z_{1-\beta}, T_n^{\text{Jac}}(\hat{F}_n) - n^{-1/2}S_n^{\text{Jac}}(\alpha_n^{\text{Jac}}(\hat{F}_n), \hat{F}_n)z_{\beta}]. \tag{3.6}
\]
This choice is asymptotically justified by theorem 3.1, lemma 3.1, and lemma 4.14. On the other hand, there is a major flaw with this choice. This is exactly the same choice that would be used if we had decided \textit{a priori} to use a trimmed mean with fixed trimming proportion equal to \( \alpha_n^{\text{Jac}}(\hat{F}_n) \). Therefore that approximation does not account for the fact that the trimming proportion is chosen adaptively. It is intuitively clear that since \( \alpha_n^{\text{Jac}}(\hat{F}_n) \) chooses the trimming proportion for which \( S_n^{\text{Jac}}(\alpha, \hat{F}_n) \) is smallest, the estimate of variance for \( T_n^{\text{Jac}}(\hat{F}_n) \) should be larger than \( S_n^{\text{Jac}}(\alpha, \hat{F}_n) \).

Another approximation to \( J_n^{\text{Jac}}(x, F) \) is that given by the bootstrap distribution \( J_n^{\text{Jac}}(x, \hat{F}_n) \). In order to get a better understanding of what is meant by \( J_n^{\text{Jac}}(x, \hat{F}_n) \), we need the following notation and conventions.

Let \( X_1, X_2, \ldots, X_n \) be distributed according to \( F \), with corresponding empirical distribution function \( \hat{F}_n \). Conditionally on having observed \( \hat{F}_n \), we let \( Y_1, Y_2, \ldots, Y_n \) be distributed according to \( \hat{F}_n \), with corresponding empirical distribution function \( \hat{G}_n \). This convention differs from the usual convention introduced in Efron (1979). That convention consists of using \( X_1^*, X_2^*, \ldots, X_n^* \) instead of \( Y_1, Y_2, \ldots, Y_n \), and of putting a * on bootstrap probability statements to differentiate them from probability statements under the distribution \( F \). The reason for using this convention rather than the more established one is that we will be considering a double bootstrap distribution which would therefore require looking at \( X_1^{**}, X_2^{**}, \ldots, X_n^{**} \). Such notation is not very attractive. Instead, as you might have guessed, we will let \( Z_1, Z_2, \ldots, Z_n \) be distributed according to \( \hat{G}_n \) (conditionally), with empirical distribution function \( \hat{H}_n \). The idea is simple: \( X, Y, \) and \( Z \) stand for the random variables for the original, bootstrap, and double bootstrap distributions, respectively, and \( \hat{F}_n, \hat{G}_n, \) and \( \hat{H}_n \) for the corresponding empirical distribution functions.

Therefore, we have

\[
J_n^{\text{Jac}}(x, \hat{F}_n) = \text{Prob}_{\hat{F}_n}\{n^{1/2}[T(\alpha_n^{\text{Jac}}(\hat{G}_n), \hat{G}_n) - T(\alpha_n^{\text{Jac}}(\hat{F}_n), \hat{F}_n)] \leq x\}
\]

\[
= \text{Prob}_{\hat{F}_n}\{n^{1/2}[T_n^{\text{Jac}}(\hat{G}_n) - T_n^{\text{Jac}}(\hat{F}_n)] \leq x\}.
\]
So, as in equation (3.4), we can write

\[ 1 - 2\beta = \text{Prob}_F \{ x_{\beta}^{\text{Jac}}(F) \leq n^{1/2} [ T_n^{\text{Jac}}(\hat{F}_n) - \theta ] \leq x_{1-\beta}^{\text{Jac}}(F) \} \]

\[ = \text{Prob}_F \{ x_{\beta}^{\text{Jac}}(\hat{F}_n) \leq n^{1/2} [ T_n^{\text{Jac}}(\hat{G}_n) - T_n^{\text{Jac}}(\hat{F}_n) ] \leq x_{1-\beta}^{\text{Jac}}(\hat{F}_n) \}. \]

The approximate confidence interval consists, once again, of inverting the probability statement (3.4) and using \( x_{\beta}^{\text{Jac}}(\hat{F}_n) \) instead of \( x_{\beta}^{\text{Jac}}(F) \). The approximate \( 1 - 2\beta \%) \) bootstrap two-sided confidence interval is then given by:

\[ \theta \in [ T_n^{\text{Jac}}(\hat{F}_n) - n^{-1/2} x_{1-\beta}^{\text{Jac}}(\hat{F}_n), \ T_n^{\text{Jac}}(\hat{F}_n) - n^{-1/2} x_{\beta}^{\text{Jac}}(\hat{F}_n) ]. \] (3.7)

Note that this is not the percentile bootstrap confidence interval which, in this notation, can be written as:

\[ \theta \in [ T_n^{\text{Jac}}(\hat{F}_n) + n^{-1/2} x_{1-\beta}^{\text{Jac}}(\hat{F}_n), \ T_n^{\text{Jac}}(\hat{F}_n) + n^{-1/2} x_{\beta}^{\text{Jac}}(\hat{F}_n) ]. \] (3.8)

Look for instance at the lower bound. In (3.7), we subtract a multiple of the \( \beta \) quantile of \( J_n^{\text{Jac}}(x, \hat{F}_n) \) from the estimate, whereas in (3.8) we subtract a multiple of the \( 1 - \beta \) quantile. Hall (1988a) referred to this phenomenon as “looking in a table backwards”. Provided that \( J_n^{\text{Jac}}(\cdot, \hat{F}_n) \) is symmetric about \( T_n^{\text{Jac}}(\hat{F}_n) \), then the two intervals are clearly identical. This should nearly be the case since \( J_n^{\text{Jac}}(\cdot, F) \) is symmetric about \( T(a_n^{\text{Jac}}(F), F) = \theta \). The bootstrap interval given by (3.7) has yet to receive a definitive name, although Tibshirani (1984) referred to it as the pivotal bootstrap, and Hall (1988a) called it the hybrid method.

This approximation is very different from the previous one. In the first case, we consider the class of normal distributions \( N(0, \cdot) \) and we select the “closest” one according to an estimate of the asymptotic variance of the estimator. This class has nothing to do with \( J_n^{\text{Jac}}(x, F) \), apart from the fact that, asymptotically, it converges to a member of the class. On the other hand, the approximation based on the bootstrap distribution considers the class of distributions \( J_n^{\text{Jac}}(x, \cdot) \), and selects the one which is indexed by our best guess to \( F \), namely \( \hat{F}_n \). In this case, \( J_n^{\text{Jac}}(x, F) \) is a member of the class. Note that the family of distributions \( J_n^{\text{Jac}}(x, \cdot) \) explicitly account for the adaptive nature of \( T_n^{\text{Jac}}(\hat{F}_n) \). Philosophically, the bootstrap approximation seems more justifiable. But is
it asymptotically justified, as is the normal approximation? We will find out the answer shortly, but first let’s see how one actually computes such a bootstrap confidence interval.

In this problem, as in most problems where the bootstrap is used in practice, the bootstrap quantiles \( x_{\beta}^{\text{Jae}}(\hat{F}_n) \) cannot be expressed in closed form. We must therefore resort to a Monte Carlo simulation to find an estimate \( \hat{x}_{\beta}^{\text{Jae}}(\hat{F}_n) \). So \( \hat{x}_{\beta}^{\text{Jae}}(\hat{F}_n) \) is an estimate of \( x_{\beta}^{\text{Jae}}(\hat{F}_n) \) which is itself an estimate for \( x_{\beta}^{\text{Jae}}(F) \). The simulation is done as follows.

- Let \( X_1, X_2, \ldots, X_n \) be a sample of size \( n \) from \( F \) with empirical distribution function \( \hat{F}_n \).
- For \( i = 1, \ldots, B \) do:
  - Let \( Y_1^i, Y_2^i, \ldots, Y_n^i \) be i.i.d. from \( \hat{F}_n \), with empirical distribution function \( \hat{G}_n^i \).
  - Compute \( T_i = n^{1/2}[T(\hat{\alpha}_n^{\text{Jae}}(\hat{G}_n^i), \hat{G}_n^i) - T(\hat{\alpha}_n^{\text{Jae}}(\hat{F}_n), \hat{F}_n)] \)
  - Let \( T_{[1]}, T_{[2]}, \ldots, T_{[B]} \) be the ordered \( T \)'s. Let \( \hat{x}_{\beta}^{\text{Jae}}(\hat{F}_n) = T_{[\beta B]} \) and \( \hat{x}_{1-\beta}^{\text{Jae}}(\hat{F}_n) = T_{[(1-\beta)B]} \)
  - The interval is \( [T(\hat{\alpha}_n^{\text{Jae}}(\hat{F}_n), \hat{F}_n) - n^{-1/2}\hat{x}_{1-\beta}^{\text{Jae}}(\hat{F}_n), T(\hat{\alpha}_n^{\text{Jae}}(\hat{F}_n), \hat{F}_n) - n^{-1/2}\hat{x}_{\beta}^{\text{Jae}}(\hat{F}_n)] \)

This was an illustration of how, in principle, bootstrap endpoints should be computed. In practice, one would not exactly follow that algorithm. The \( T_i \)'s involve the multiplication and subtraction of two constants independent of \( i \). So a better way to implement this algorithm consists of replacing the formula for \( T_i \) by \( T_i = T(\hat{\alpha}_n^{\text{Jae}}(\hat{G}_n^i), \hat{G}_n^i) \). Then the interval becomes

\[
\theta \in [2T(\hat{\alpha}_n^{\text{Jae}}(\hat{F}_n), \hat{F}_n) - T_{[(1-\beta)B]}, 2T(\hat{\alpha}_n^{\text{Jae}}(\hat{F}_n), \hat{F}_n) - T_{[\beta B]}].
\]

Let’s now state the result establishing the asymptotic validity of the bootstrap approximation.
**Theorem 3.2**: Let $F$ satisfy assumptions (A.1) through (A.3). Let $n^{1/4}b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let's further assume (A.5). Then

$$\sup_x |J_n^{Jae}(x, F) - J_n^{Jae}(x, \hat{F}_n)| \to 0 \quad \text{in probability},$$

as $n \to \infty$.

Basically, this result says that with high probability $J_n^{Jae}(x, F)$ and $J_n^{Jae}(x, \hat{F}_n)$ will be uniformly close when $n$ is large. This result is about the distribution functions. It is easy to see that this implies that for fixed $\beta$, the two corresponding quantiles, $x^{Jae}_\beta(F)$ and $x^{Jae}_\beta(\hat{F}_n)$, are also asymptotically close which is sufficient to get the following result:

**Corollary 3.1**: Let $F$ satisfy assumption (A.1) through (A.3). Let $n^{1/4}b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let's further assume (A.5). Then the bootstrap confidence interval given by (3.7) has an asymptotic coverage probability of $1 - 2\beta$.

Therefore, both the standard confidence interval (3.6) and the bootstrap interval (3.7) have asymptotically the claimed coverage probability. So they are both said to be consistent. But this is not sufficient to conclude that one method is superior to the other, either asymptotically or for finite samples. Second order results would be rather difficult to obtain because of the adaptive choice of the trimming proportion. In any case, Hall (1988a) showed that for a certain class of statistics whose distribution admits an Edgeworth expansion, both the standard interval and this hybrid bootstrap interval are only first order correct. A notable exception is in the case of a confidence interval for a slope in regression where, because of the underlying symmetry in the model, the hybrid method is second order correct whereas the standard method remains first order correct (Hall (1988b)). Hence it is very likely that in this case the two methods would only be first order correct.

Second order correct methods usually involve more work. For instance, the BC$_a$ interval of Efron (1987) requires the calculation of a so-called acceleration constant which would undoubtedly be difficult to compute in this adaptive setting. On the other hand,
the percentile-t as discussed, for instance, in Hall (1988a) could, in principle, be computed. Instead of bootstrapping the root $T_n^{Jae}(\hat{F}_n) - T_n^{Jae}(F)$, the percentile-t consists of bootstrapping a "studentized" root $(T_n^{Jae}(\hat{F}_n) - T_n^{Jae}(F))/\sigma(\alpha_n^{Jae}(\hat{F}_n), \hat{F}_n)$. Therefore, for each bootstrap sample, one computes an estimate of the asymptotic variance of Jaeckel's estimator. Note that this estimate is readily available, having been computed in order to determine $\alpha_n^{Jae}$. Letting $K_n^{Jae}(\cdot, F)$ be the distribution function of $n^{1/2}[T_n^{Jae}(\hat{F}_n) - T_n^{Jae}(F)]/\sigma(\alpha_n^{Jae}(\hat{F}_n), \hat{F}_n)$, and $y_n^{Jae}(F) = K_n^{Jae-1}(\beta, F)$, the percentile-t interval is given by

$$\theta \in [T_n^{Jae}(\hat{F}_n) - n^{-1/2}y_n^{Jae}(\hat{F}_n)\sigma(\alpha_n^{Jae}(\hat{F}_n), \hat{F}_n), T_n^{Jae}(\hat{F}_n) - n^{-1/2}y_n^{Jae}(\hat{F}_n)\sigma(\alpha_n^{Jae}(\hat{F}_n), \hat{F}_n)].$$

(3.9)

But even though this interval can be computed, it need not be second order correct as the sufficient conditions of Hall (1988a) are not satisfied. Another second order correct method, the automatic percentile, due to DiCiccio and Romano (1987, 1988a,b) also requires the same type of conditions.

Basically, the condition needed by Hall and DiCiccio and Romano is that the estimator admits an Edgeworth expansion. Whereas the fixed trimming proportion trimmed mean has an Edgeworth expansion, as was shown by Bjerve (1974) in his dissertation, it is unclear whether such a result holds for an adaptive trimmed mean. In any case, the performance of different confidence intervals for small sample sizes will be assessed in chapter 5.

Finally, note that the bootstrap methodology was not available when Jaeckel worked on this problem. In fact, the bootstrap was introduced by Efron (1979), and bootstrap confidence intervals, even though mentioned in 1979, became a centerpiece of the bootstrap methodology in Efron (1981). Consequently, before 1979, one had to settle for an interval based on asymptotic normality in order to compute a confidence interval for the location parameter based on Jaeckel's estimator.
3.3. Bootstrap Adaptive Trimmed Mean

We have just seen that Jaeckel's estimator is a trimmed mean with a trimming proportion that minimizes estimates of the asymptotic variance of the $\alpha$-trimmed mean. We have seen that in so doing, its asymptotic variance is as small as possible. But these results obviously are of no comfort when you have a sample of size 10, especially when the trimming proportion is selected so as to minimize estimates of the asymptotic variance. The trimming proportion that minimizes the finite sample variance might be rather different than the one which minimizes the asymptotic variance. Or is it?

Figure 3.1 shows a plot of the asymptotic variance and the suitably standardized variance of $\alpha$-trimmed means for samples of size 10 and 20 from the standard Gaussian distribution as a function of the trimming proportion $\alpha$. Due to their continuity, I used the trimmed means of definition 2.2 instead of definition 2.1. But since our estimators only consider values of $\alpha$ such that $n\alpha$ is an integer, it does not matter because both definitions agree for those values. Notice the knots. They occur precisely at those values of $\alpha$ at which the asymptotic variance is (possibly, depending on $\alpha_0$ and $b_n$) estimated. It is clear that for the normal distribution, it does not matter that we are estimating the asymptotic variance rather than the finite sample variance as the curves are almost identical at those values.

In Figure 3.2, we see a similar plot for the double exponential distribution. Qualitatively, the two plots are almost opposites of one another. In this case, the asymptotic variance declines as the trimming proportion increases. Also, the shape of the finite sample variances are similar, but not all that close to the asymptotic variance. Note that the minimum asymptotic variance is obtained at the median, which is not surprising given the fact that it is the maximum likelihood estimator. But note that the median does not minimize the finite sample variance, as its curve bends slightly upward close to the median (recall that the median is a trimmed mean with trimming proportion equal to 0.4 and 0.45, respectively for 10 and 20 observations). So in the case of the double exponential distribution, it might be worth minimizing a characteristic of the finite sample distribution of
Figure 3.1: Standardized Variance of Trimmed Means
Normal Distribution

Figure 3.2: Standardized Variance of Trimmed Means
Double Exponential Distribution
the trimmed means rather than the asymptotic variance.

In light of these plots, we set the following two goals for our alternative adaptive estimators.

GOALS:

- Get the smallest possible asymptotic variance among the class of estimators considered.

- Get good finite sample properties as compared to the estimators in the class.

We will now introduce two bootstrap adaptive trimmed means that will attempt to achieve these two goals. In both cases, their trimming proportion will be chosen so as to minimize bootstrap estimates of a characteristic of the finite sample distribution of the trimmed means. One of them will minimize the finite sample variance while the other one will minimize the interquartile range of the finite sample distribution.

The idea of a bootstrap adaptive trimmed mean is not new. De Jongh and de Wet (1985) considered an adaptive trimmed regression estimator. But they only reported the results of a finite sample Monte Carlo simulation; they did not study its asymptotic behavior. Also, Hsieh and Manski (1987), in a Monte Carlo study of the performance of adaptive maximum likelihood estimators, considered an estimator whose tuning parameter (the smoothing parameter in the kernel used for the density estimation) was adaptively selected by minimizing bootstrap estimates of the mean square error.

Again, let's consider the setup introduced in section 2.3 of chapter 2. Let's first define

\[ S_n^{\text{Var}}(\alpha, F) = \text{Variance of } M_n(\alpha, \hat{F}_n, F), \]

and,

\[ S_n^{\text{IQR}}(\alpha, F) = \text{IQR of } M_n(\alpha, \hat{F}_n, F), \]

where IQR stands for the interquartile range, i.e., the difference between the upper and
lower quartiles. Also, let

\[ \alpha_n^{Var}(\hat{F}_n) = \arg \min_{\alpha \in A_n} S_n^{Var}(\alpha, \hat{F}_n), \]

and

\[ \alpha_n^{IQR}(\hat{F}_n) = \arg \min_{\alpha \in A_n} S_n^{IQR}(\alpha, \hat{F}_n), \]

where \( A_n \) is defined in (3.2). So, \( \alpha_n^{Var}(F) \) chooses the trimming proportion that minimizes the variance of the finite sample distribution of \( T(\alpha, \hat{F}_n) \), whereas \( \alpha_n^{IQR}(F) \) minimizes its interquartile range. Minimizing the finite sample variance is a natural alternative to minimizing the asymptotic variance as Jaeckel’s estimator does. But this method might not work for other classes of estimators. As we shall see in the next chapter, in order to get similar results to Jaeckel’s estimator, \( S_n^{Var}(\alpha, F) \) must be shown to converge to \( \sigma^2(\alpha, F) \), uniformly in \( \alpha \). As we have seen in chapter 2, \( M_n(\alpha, \hat{F}_n, F) \) converges weakly. But, as is well known, this is not sufficient to imply the convergence of the moments. On the other hand, weak convergence of the distribution functions is equivalent to weak convergence of the quantile function. Thus, since the interquartile range is the difference between the upper and lower quartiles, if the quantile function of the asymptotic distribution is continuous at .75 and .25, then the interquartile range of the finite sample distribution will automatically converge to the interquartile range of the asymptotic distribution. Of course, one must still show that this convergence is uniform over the class of estimators, but at least the pointwise convergence is guaranteed. Moreover, if the asymptotic distribution is a scale family (as is the case here), then the value of \( \alpha \) which minimizes the asymptotic variance is equal to that which minimizes the asymptotic interquartile range. So if such an estimator works, it will also minimize the asymptotic variance as does Jaeckel’s estimator.

Let’s denote the bootstrap adaptive trimmed mean estimate of location based on the finite sample variance by \( T_n^{Var}(\hat{F}_n) = T(\alpha_n^{Var}(\hat{F}_n), \hat{F}_n) \) and let \( J_n^{Var}(x, F) \) be the distribution function of \( M_n(\alpha_n^{Var}(\hat{F}_n), \hat{F}_n, F) \), i.e.,

\[ J_n^{Var}(x, F) = \text{Prob}_F\{n^{1/2}[T_n^{Var}(\hat{F}_n) - T(\alpha_n^{Var}(F), F)] \leq x\} \]

\[ = \text{Prob}_F\{n^{1/2}[T_n^{Var}(\hat{F}_n) - \theta] \leq x\}, \]
where the second equality is valid whenever $F$ is symmetric about $\theta$. Similarly, let's denote the bootstrap adaptive trimmed mean estimate of location based on the finite sample interquartile range by $T_n^{lqr}(\hat{F}_n) = T(\alpha_n^{lqr}(\hat{F}_n), \hat{F}_n)$ and let $J_n^{lqr}(x, F)$ be the distribution function of $M_n(\alpha_n^{lqr}(\hat{F}_n), \hat{F}_n, F)$, i.e.,

$$ J_n^{lqr}(x, F) = \text{Prob}_F\{n^{1/2}[T_n^{lqr}(\hat{F}_n) - T(\alpha_n^{lqr}(F), F)] \leq x \} $$

$$ = \text{Prob}_F\{n^{1/2}[T_n^{lqr}(\hat{F}_n) - \theta] \leq x \}, $$

with the same provision as before about the second equality.

Obviously, these estimators are much more computer intensive than was Jaeckel's estimator. Since, neither $S_n^{\text{Var}}(\alpha, \hat{F}_n)$ nor $S_n^{lqr}(\alpha, \hat{F}_n)$ can be expressed in closed form, they must be estimated by Monte Carlo sampling. Now let's take a look at the properties of this estimator.

**Lemma 3.2:** Let $F$ satisfy assumptions (A.1) through (A.3). Let $n^{1/2}b_n \to \infty$ as $n \to \infty$. Then

$$ \alpha_n^{\text{Var}}(\hat{F}_n) \to A_0(F) \quad \text{in probability}. $$

As we have seen in Jaeckel's case, this result is one of the key in proving the following theorem.

**Theorem 3.3:** Let $F$ satisfy assumptions (A.1) through (A.3). Let $n^{1/2}b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let's further assume (A.5). Then

$$ \sup_x |J_n^{\text{Var}}(x, F) - \Phi(x\sigma(A_0(F), F))| \to 0 $$

as $n \to \infty$ where $\Phi(x)$ is the distribution function of the standard normal distribution, and $\sigma^2(\alpha, F)$ is given by (2.2).

The corresponding results for the bootstrap adaptive trimmed mean based on the interquartile range are as follows:
Lemma 3.3: Let $F$ satisfy assumptions (A.1) through (A.3). Let $n^{1/2}b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let’s further assume (A.4).

$$\alpha_n^{iqr}(\hat{F}_n) \to A_0(F) \quad \text{in probability.}$$

Theorem 3.4: Let $F$ satisfy assumptions (A.1) through (A.3). Let $n^{1/2}b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let’s further assume (A.4) and (A.5). Then

$$\sup_x |J_n^{iqr}(x, F) - \Phi(x\sigma(A_0(F), F))| \to 0$$

as $n \to \infty$ where $\Phi(x)$ is the distribution function of the standard normal distribution, and $\sigma^2(\alpha, F)$ is given by (2.2).

Therefore, as for Jaeckel’s estimator, both of these bootstrap adaptive trimmed means have an asymptotic variance equal to the smallest one among trimmed means with fixed trimming proportion $\alpha$ for either $\alpha \in [\alpha_0, 1/2]$ or $\alpha \in [\alpha_0, \alpha_1]$ depending on $\{b_n\}$. See assumption A. Hence both of these estimators do asymptotically as well as possible within the class of trimmed means as does Jaeckel’s estimator. Thus both $T_n^{Var}(\hat{F}_n)$ and $T_n^{iqr}(\hat{F}_n)$ achieve the first goal previously stated. We will look at the second goal in chapter 5.

Looking at the different assumptions, note that assumption (A.5) is not needed in lemmas 3.1 through 3.3, even though it is required in theorems 3.1 through 3.4. Hence it is not clear what happens to the distribution function of the adaptive trimmed means when it is not met, although it is known that the random trimming proportion will converge in probability to the trimming proportion which minimizes the asymptotic variance. Also, note that the bootstrap adaptive trimmed mean based on the interquartile range further requires that the density at the median be bounded whenever $b_n \to 0$.

Note that there is a small difference between theorem 3.1 and theorems 3.3 and 3.4: the rate at which the largest possible trimming proportion considered in the definitions of $\alpha_n^{jae}$, $\alpha_n^{Var}$, and $\alpha_n^{iqr}$ can approach 1/2 differ. In Jaeckel’s case, we need $n^{1/4}b_n \to \infty$ which is slower than $n^{1/2}b_n \to \infty$ in the bootstrap adaptive estimators. This is of course a minor point.
Another interesting point to note is that the interquartile range of the distribution of a centralized statistic is equal to the length of a 50% confidence interval based on that statistic. Therefore, \( T_n^{1qr}(\hat{F}_n) \) selects the trimming proportion which minimizes the length of 50% bootstrap confidence intervals. Obviously, there is nothing magic about the confidence level 50%. For instance, we might want to use higher confidence levels such as 80% or 90%. All the results in this section and the next one would remain valid. But the small sample results of chapter 5 could have been quantitatively different. Let's now look at confidence intervals for the location parameter based on that bootstrap estimator.

### 3.4. Confidence Intervals Based on a Bootstrap Adaptive Trimmed Mean

Once again, two alternative methods to construct approximate confidence intervals for \( \theta \) based on the two bootstrap adaptive trimmed means are available. The first one is based on the asymptotic normality of the proposed estimator. As motivated in section 3.2, the approximate \( 1 - 2\beta \)% two-sided standard confidence intervals are given by

\[
\theta \in [T_n^{Var}(\hat{F}_n) - n^{-1/2} S^{Jae}(\alpha_n^{Var}(\hat{F}_n), \hat{F}_n) z_{1-\beta}, T_n^{Var}(\hat{F}_n) - n^{-1/2} S^{Jae}(\alpha_n^{Var}(\hat{F}_n), \hat{F}_n) z_\beta],
\]

(3.10)

and

\[
\theta \in [T_n^{1qr}(\hat{F}_n) - n^{-1/2} S^{Jae}(\alpha_n^{1qr}(\hat{F}_n), \hat{F}_n) z_{1-\beta}, T_n^{1qr}(\hat{F}_n) - n^{-1/2} S^{Jae}(\alpha_n^{1qr}(\hat{F}_n), \hat{F}_n) z_\beta].
\]

(3.11)

These intervals will asymptotically have the right coverage probability in light of lemmas 3.2 and 3.3, lemma 4.14 and theorems 3.3 and 3.4. But as with Jaeckel’s estimator, these intervals are identical to the one that would be obtained if the trimming proportion had been fixed \textit{a priori} at \( \alpha_n^{Var}(\hat{F}_n) \) and \( \alpha_n^{1qr}(\hat{F}_n) \), respectively.

The other possibility is to use bootstrap intervals. Note that such intervals involve a double bootstrap as the estimators themselves include a bootstrap step. They are therefore highly computer intensive. So as was motivated earlier, the \( 1 - 2\beta \)% two-sided bootstrap
confidence interval based on the bootstrap adaptive trimmed means are given by

$$\theta \in [T_{n}^{\text{Var}}(\hat{F}_n) - n^{-1/2} x_{1-\beta}^{\text{Var}}(\hat{F}_n), T_{n}^{\text{Var}}(\hat{F}_n) - n^{-1/2} x_{\beta}^{\text{Var}}(\hat{F}_n)],$$  \hspace{1cm} (3.12)$$

and$$\theta \in [T_{n}^{\text{Iqr}}(\hat{F}_n) - n^{-1/2} x_{1-\beta}^{\text{Iqr}}(\hat{F}_n), T_{n}^{\text{Iqr}}(\hat{F}_n) - n^{-1/2} x_{\beta}^{\text{Iqr}}(\hat{F}_n)],$$  \hspace{1cm} (3.13)$$

where $x_{\beta}^{\text{Var}}(F) = J_{n}^{\text{Var}}(\beta, F)$ and $x_{\beta}^{\text{Iqr}}(F) = J_{n}^{\text{Iqr}}(\beta, F)$ are the $\beta^{th}$ percentile of the distribution of $M_{n}(\alpha_{n}^{\text{Var}}(\hat{F}_n), \hat{F}_n, F)$ and $M_{n}(\alpha_{n}^{\text{Iqr}}(\hat{F}_n), \hat{F}_n, F)$, respectively.

As an illustration, let’s now see how the interval (3.13) is computed.

- Let $X_1, X_2, \ldots, X_n$ be a sample of size $n$ from $F$ with empirical distribution function $\hat{F}_n$.
- For $i = 1, \ldots, B_1$ do:
  - Let $Y_1^i, Y_2^i, \ldots, Y_n^i$ be i.i.d. from $\hat{F}_n$, with empirical distribution function $\hat{G}_n^i$.
  - For $j = 1, \ldots, B_2$ do:
    - Let $Z_1^{i,j}, Z_2^{i,j}, \ldots, Z_n^{i,j}$ be i.i.d. from $\hat{G}_n^i$, with empirical distribution function $\hat{H}_n^{i,j}$.
    - For each $\alpha \in A_n$ do:
      - Compute $T_j^{i}(\alpha) = T(\alpha, \hat{H}_n^{i,j})$
    - For each $\alpha \in A_n$ do:
      - Let $\hat{S}_n^{\text{Iqr}}(\alpha, \hat{G}_n^i) = T_{(0.75B_2)}^{i}(\alpha) - T_{(0.25B_2)}^{i}(\alpha)$ where $T_{(1)}^{i}(\alpha), T_{(2)}^{i}(\alpha), \ldots, T_{(B_2)}^{i}(\alpha)$ are the ordered $T$’s.
      - Let $\hat{a}_n^{\text{Iqr}}(\hat{G}_n^i) = \min_{\alpha \in A_n} \hat{S}_n^{\text{Iqr}}(\alpha, \hat{G}_n^i)$
      - Let $D_i = T(\hat{a}_n^{\text{Iqr}}(\hat{G}_n^i), \hat{G}_n^i) - T(\hat{a}_n^{\text{Iqr}}(\hat{F}_n), \hat{F}_n)$
      - Let $D_{(1)}, D_{(2)}, \ldots, D_{(B_1)}$ be the ordered $D$’s. Let $\hat{z}_n^{\text{Iqr}}(\hat{F}_n) = D_{(\beta B_1)}$ and $\hat{z}_n^{\text{Iqr}}(\hat{F}_n) = D_{((1-\beta)B_1)}$
    - The interval is $[T(\hat{a}_n^{\text{Iqr}}(\hat{F}_n), \hat{F}_n) - \hat{z}_n^{\text{Iqr}}(\hat{F}_n), T(\hat{a}_n^{\text{Iqr}}(\hat{F}_n), \hat{F}_n) - \hat{z}_n^{\text{Iqr}}(\hat{F}_n)]$
Note that we put hats (') on every $S_n^\text{lqr}$, $\alpha_n^\text{lqr}$, and $x_\beta^\text{lqr}$ to stress the fact that they really are (Monte Carlo) estimates of those values. As was mentioned before, and have clearly illustrated in the above algorithm, this confidence interval involves a double bootstrap. It requires $B_1 B_2$ bootstrap samples of size $n$ that must each be ordered, a procedure which itself requires $O(n \log n)$ operations. It must be stressed that $B_2$ need not be as large as $B_1$ (which is usually of the order of 1000) because those bootstrap samples are used in the selection of the trimming proportion $\alpha$. We do not necessarily need a good estimate of the interquartile range, but only a good estimate of the value of $\alpha$ which minimizes those estimates of the interquartile range. So we are only interested in the ordering of the different estimates. In particular, it wouldn’t matter if those estimates were biased, as long as the bias is constant over $\alpha$.

The following results establish the asymptotic validity of the bootstrap confidence intervals.

**Theorem 3.5:** Let $F$ satisfy assumptions (A.1) through (A.3). Let $n^{1/2} b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let’s further assume (A.5). Then

$$\sup_x |J_n^{\text{Var}}(x, F) - J_n^{\text{Var}}(x, \hat{F}_n)| \to 0 \quad \text{in probability,}$$

as $n \to \infty$.

**Theorem 3.6:** Let $F$ satisfy assumptions (A.1) through (A.3). Let $n^{1/2} b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let’s further assume (A.4) and (A.5). Then

$$\sup_x |J_n^{\text{lqr}}(x, F) - J_n^{\text{lqr}}(x, \hat{F}_n)| \to 0 \quad \text{in probability,}$$

as $n \to \infty$.

As we have seen in Jaeckel's case, this is sufficient to imply the following corollaries.

**Corollary 3.2:** Let $F$ satisfy assumption (A.1) through (A.3). Let $n^{1/2} b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let’s further assume (A.5). Then the bootstrap confidence interval given by (3.12) has an asymptotic coverage probability of $1 - 2\beta$. 

Section 3.4: Confidence Intervals Based on a Bootstrap Adaptive Trimmed Mean 39
Corollary 3.3: Let $F$ satisfy assumption (A.1) through (A.3). Let $n^{1/2}b_n \to \infty$ as $n \to \infty$. If $b_n \to 0$, let's further assume (A.4) and (A.5). Then the bootstrap confidence interval given by (3.13) has an asymptotic coverage probability of $1 - 2\beta$.

In the next section, two alternative estimators of the distribution $F$ will be considered, namely, a kernel distribution function estimator and a symmetric estimator.

3.5. Estimators Based on Alternative Estimators of $F$

A bootstrap estimate of a functional evaluated at $F$ is the value of that functional replacing the distribution function $F$ by an estimate $\hat{F}_n$. In the previous sections, the empirical distribution function $\hat{F}_n$ was used. This certainly is not a bad choice. It is asymptotically minimax and is the nonparametric maximum likelihood estimator of $F$. See Beran (1984). On the other hand, it is discrete even though, by assumption, $F$ is continuous, at least on $[F^{-1}(\alpha_0), F^{-1}(1 - \alpha_0)]$. Moreover, unlike $F$, the empirical distribution function is not symmetric.

As alternatives to using the empirical distribution function as the estimator of $F$ in the bootstrap steps of the estimators $T_{nk}^\text{Var}$ and $T_{nk}^\text{Jae}$, or in bootstrap confidence intervals based either on them or on $T_{nk}^\text{Jae}$, a continuous estimator $\hat{F}_{n,c_n}$ and a symmetric estimator $\hat{F}_n$ will be considered. Let's start with $\hat{F}_{n,c_n}$.

Let $K$ be a distribution function with density $k$. Let

$$\hat{F}_{n,c_n}(x) = \frac{1}{n} \sum_{i=1}^{n} K \left( \frac{x - X_i}{c_n} \right),$$

(3.14)

where $c_n$ is a sequence of smoothing parameters which converges to 0 as $n$ tends to infinity. We call $\hat{F}_{n,c_n}$ the distribution function estimate of $F$ based on kernel $k$. Because of its continuity, a bootstrap estimator based on such an estimate of $F$ is called a smooth bootstrap.

Differentiating (3.14) gives the well-known kernel density estimate introduced by Rosenblatt (1956) and Parzen (1962). Nadaraya (1965) studied the distribution function estimate $\hat{F}_{n,c_n}$ and showed that, under mild conditions, it has asymptotically the same
mean and variance as \( \hat{F}_n \). Azzalini (1981) has obtained second order results that show a possible asymptotic improvement in the estimation of \( F(x) \) by using \( \hat{F}_{n,c_n}(x) \) instead of \( \hat{F}_n(x) \) provided certain regularity conditions on \( F \) are met and that the sequence \( c_n \) converges to 0 at a certain rate.

The use of an estimator like \( \hat{F}_{n,c_n} \) in the bootstrap, or a variation thereof, was first suggested in the original paper of Efron (1979). He recognized that for certain statistics, the discreteness of the bootstrap distribution based on the empirical distribution function might lead to some problems. The question of whether the smooth bootstrap is superior to the usual bootstrap, and for which smoothing parameters \( c_n \) has been worked on by a few people although, in most cases, no definitive answers exist, especially in small samples. Silverman and Young (1987) have shown that in estimating a functional \( T(F) \), there exists an unknown smoothing parameter \( c \) for which the (finite sample) mean squared error of the smooth bootstrap estimate \( T(\hat{F}_{n,c_n}) \) is smaller than that of the usual bootstrap estimate \( T(\hat{F}_n) \), provided that certain conditions on both \( T \) and the unknown distribution function \( F \) (!) are satisfied. Note that this result does not specify the value of the smoothing parameter \( c \) for which the smooth bootstrap is superior. So this result lacks a certain usefulness in practice. Another paper on the subject was written by Hall, DiCiccio, and Romano (1988). They show that in estimating the variance of a sample quantile, the rate of convergence of the relative error can be improved by using a smooth bootstrap instead of the usual bootstrap. Finally, Romano (1988) has showed that if the functional of interest can be viewed as a functional of the density, bootstrap methods can even be inconsistent unless the resampling is done from a sufficiently smooth estimate of the distribution function. There is still a lot of work to be done in this area, especially for small samples. A result in the spirit of the work of Silverman and Young (1987) which would give an explicit value of the soothing parameter \( c \) would certainly be welcomed.

From a computational point of view, it is easy to see that a bootstrap sample from \( \hat{F}_{n,c_n} \) can be obtained by choosing among \( X_1, X_2, \ldots, X_n \) with replacement and adding an independent auxiliary random variable with density \( c_n^{-1}f(x/c_n) \). So, as long as a good
random number generator for the density $f$ is available, the smooth bootstrap is just as easy to implement as the ordinary bootstrap based on the empirical distribution function.

The following theorem tells us that theorems 3.1 through 3.6 remain valid if we use $\hat{F}_{n,c_n}$ instead of $\hat{F}_n$, provided that certain conditions on the distribution function $K$ are satisfied.

**THEOREM 3.7:** Suppose that $F$ has a bounded density and that the (integrated) kernel $K$ has a finite first moment. If either $n^{1/2}c_n \to 0$ as $n \to \infty$ or $n^{1/2}c_n^2 \to 0$ as $n \to \infty$ and that it is further assumed that $f$ is differentiable with a uniformly bounded derivative and $K$ has a finite second moment, then theorems 3.1 through 3.6 remain valid if $\hat{F}_n$ is replaced by the smoother $\hat{F}_{n,c_n}$.

Note that $K$ need not have a density, although this might render the application of the method unnecessarily complicated. On the other hand, the density of $F$ must be at least bounded. The rate at which the smoothing parameter $c_n$ may decrease to 0 depends on the assumed smoothness of $F$. It is interesting to note that depending whether the goal is to estimate the density or the distribution function, the "best" value of the smoothing parameter is different. For the problem of density estimation, the optimal choice for minimizing the integrated mean square error is $c_n = O(n^{-1/5})$, (see e.g. Tapia and Thompson (1978)), whereas the optimal choice is $c_n = O(n^{-1/3})$ for the problem of estimating the distribution function (see e.g. Azzalini (1981)). Note that the optimal rate for density estimation cannot be achieved here, whereas the optimal rate for the estimation of the distribution function can be achieved if extra smoothness conditions on $F$ and a finite variance for $K$ is assumed.

The other alternative estimator of the symmetric distribution function $F$ is the symmetric $\hat{F}_n$. The idea is simple: let $\hat{F}_n$ be the average of the empirical distribution function and its reflection about an estimate of the median $\hat{\theta}$. Such estimators have been studied by Schuster (1975). He showed that, under regularity conditions, $\hat{F}_n$ can be asymptotically better than $\hat{F}_n$. This is the case, for instance, if $F$ is normal, double exponential, or
Cauchy.

Once again, Efron first proposed the use of such an estimator in connection with the bootstrap in his 1979 paper. There does not seem to have been any theory published on such a symmetric bootstrap.

Specifically, let \( \theta(\hat{F}_n) = \hat{F}_n^{-1}(1/2) \) be the median of the sample. Then let

\[
\hat{F}_n(x) = (1/2)[\hat{F}_n(x) + 1 - \hat{F}_n(2\theta(\hat{F}_n) - x - 0)],
\]  

(3.15)

where \( \hat{F}_n(x - 0) = \lim_{y \to x^-} \hat{F}_n(y) \). Under weak conditions, the next theorem says that one can replace \( \hat{F}_n \) by \( \hat{F}_n \) in the previous theorems.

**THEOREM 3.8:** Suppose that \( F \) is a symmetric distribution function with a bounded density. Then theorems 3.1 through 3.6 remain valid if \( \hat{F}_n \) is replaced by the symmetric \( \hat{F}_n \).

The main reason for including theorems 3.7 and 3.8 is to illustrate the flexibility of the method of proof that will be employed in the next chapter. Bickel and Freedman (1981) mentioned that most of their asymptotic results for the bootstrap would also apply to the smooth bootstrap. But by using Beran's method of proof (Beran (1984)), one can actually make such claims explicit with very little extra efforts.

### 3.6. Asymmetric Distributions

So far, all of the results have assumed symmetry of the underlying distribution, the first half of (A.1). Such an assumption implies that every trimmed mean is estimating the same quantity, namely, the center of the distribution. In particular, an adaptive trimmed mean is also estimating the location of the distribution. In such a case, the adaptive trimmed means not only estimate the location of the distribution consistently, but their asymptotic variance are as small as possible within the class of trimmed means.

A careful look at the proofs of the results of this chapter shows that they do not depend on the symmetry of the underlying distribution, except for theorem 3.8. The only
necessary change is in the definition of \( \sigma^2(\alpha, F) \) which is given by

\[
\sigma^2(\alpha, F) = (1 - 2\alpha)^{-2} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} F(\min(s, t)) - F(s)F(t) \, ds \, dt,
\]

instead of (2.2). Note that this expression for the asymptotic variance of the trimmed mean is also valid for symmetric \( F \)'s.

Even though the results also apply when the symmetry assumption is relaxed, their interpretations differ. First, the \( \alpha \)-trimmed mean estimates a different functional for each value of \( \alpha \). So the adaptive trimmed means estimate \( T(A_\alpha(F), F) \) consistently and have the smallest asymptotic variance among the class of trimmed means. So asymptotically, we know which functional of \( F \) is being estimated, but for each finite \( n \), the adaptive trimmed mean chooses not only a trimming proportion based on the data, but also the functional being estimated.

That the asymptotic variance is minimized is certainly a good property, but the fact that the functional being estimated varies with \( n \) is more troublesome. If the asymmetry is rather mild, then the mean and the median will not be very different, and since \( T(\alpha, F) \) will be between those two values for all \( \alpha \)'s, then all trimmed means will be rather close. In such a case, the asymmetry will not play a crucial role. Most robust estimators, including the trimmed means, will be estimating more or less the same quantity. Therefore, that the adaptive trimmed means minimize the asymptotic variance among the class of trimmed means should be a valid asymptotic argument to prefer them to the fixed trimming proportion trimmed means.

On the other hand, if the distribution is so asymmetric that the mean and the median of the distribution are very far apart, the lack of a precise definition of location is more problematic. In such a case, it seems that the statistician should decide which trimmed mean of \( F \) should be estimated, and then choose the trimming proportion that minimizes bootstrap estimates of the mean squared error of the \( \alpha \)-trimmed mean in estimating the desired trimmed mean of the distribution \( F \). It is conjectured that the theory of chapter 4 would follow and that the asymptotic mean squared error of the adaptive trimmed mean
would be as small as possible within the class of trimmed means. The author hopes to pursue such an approach soon.

In conclusion, if a statistician is willing to use $\alpha$-trimmed means on asymmetric data sets to obtain a robust estimate of location, in cases where the location parameter is loosely defined, then adaptive trimmed means could also be used and will minimize asymptotic variance.

This ends this chapter in which we have presented three adaptive trimmed means which do asymptotically as well as possible within the class of trimmed means along with approximate confidence intervals that asymptotically have the claimed coverage probability. Jaeckel's estimate chooses the trimming proportion which minimizes estimates of the asymptotic variance of the trimmed mean with fixed trimming proportion $\alpha$, whereas the alternative bootstrap adaptive trimmed mean minimizes bootstrap estimates of the interquartile range and variance of their finite sample distribution. Chapter 4 contains the proofs of those results.
Chapter 4

Proofs

This chapter contains the proofs of the results in the previous chapter. In the next section, we will outline the properties of the trimmed mean process and of the three adaptive choices of the trimming proportion $\alpha$. It will become clear that this problem has a lot of structure and that other problems that share the same structure might be solved in much the same manner. The last section contains detailed proofs of the results.

4.1. Ideas of The Proofs

You might have noticed that theorems 3.1 through 3.6 have a lot in common. As it should soon become clear, there is a lot of structure in this problem. Fortunately, this structure is also shared by other similar problems. Consequently, the discussion will be more general than would be necessary to solve this problem so as to make the extensions to more complicated problems straightforward.

In chapter 3, we described adaptive procedures that involved two elements: a class of estimators indexed by a parameter $\alpha$ and a random variable $\alpha_n$. The adaptive estimator is then the randomly selected estimator indexed by $\alpha_n$. The asymptotic distribution of the adaptive estimator was given and it was noted that it is identical to the asymptotic distribution of one of the estimators in the class.

Since it is possible to construct many different adaptive estimators from the same class of estimators by defining new random choices $\alpha_n$, it might be advantageous to study the asymptotic behavior of the estimators in the class separately from the asymptotic behavior of the random variables $\alpha_n$, and then try to combine the results together. It will also be helpful and instructive to treat the class of estimators as a stochastic process. By
doing so, it will be possible to use some of the elegant results of the asymptotic theory of
stochastic processes. Such a formulation will also simplify the possible extensions to more
general problems.

So let $R_n(\alpha, \hat{F}_n, F)$ be a stochastic process. Here $\alpha$ indexes the process; for each
fixed value of $\alpha$, $R_n(\alpha, \hat{F}_n, F)$ is a random variable. In fact, each value of $\alpha$ identifies
a different root that could be used to construct an approximate confidence interval for
a given parameter of the distribution $F$. (The term root is borrowed from Beran (1987)
and refers to a "pivot" whose distribution need not be independent from $F$.) Since each
root is an explicit function of both the data and the distribution, the parameters $\hat{F}_n$
and $F$ are explicitly included in $R_n$. Even though $R_n$ is a stochastic process, its last
two parameters illustrate its specialized nature. It is constructed from independently
and identically distributed random variables from the distribution function $F$, and with
empirical distribution function $\hat{F}_n$.

The second element of an adaptive estimator is a random variable $\alpha_n(\hat{F}_n)$. We are in-
terested in computing the asymptotic distribution of the adaptive root $R_n(\alpha_n(\hat{F}_n), \hat{F}_n, F)$,
which as we indicated before should asymptotically behave like $R_n(\alpha_0, \hat{F}_n, F)$ for some
$\alpha_0$. Such a result seems to require that $R_n$ be smooth in $\alpha$ and that $\alpha_n(\hat{F}_n)$ converge to
$\alpha_0$ in some sense. Let's see just what kind of smoothness of the stochastic process and
convergence of the random variable $\alpha_n(\hat{F}_n)$ are sufficient to provide such a result.

Let's introduce the concept of $F_n$-stochastic equicontinuity as discussed in Pollard
(1984). Even though, we are only interested in the case $F_n = F$ for now, the more
general definition will be used shortly in relation to bootstrap confidence intervals. In the
following definition, $\{F_n\}$ will be a sequence of fixed distribution functions. The empirical
distribution function of a sample of size $n$ from $F_n$ will be denoted $\hat{G}_n$. Note that when
$F_n = F$, we use $\hat{F}_n$ instead of $\hat{G}_n$.

**Definition 4.1:** Let $\{F_n\}_{n=1}^{\infty}$ be an infinite sequence of distribution functions. We
will call $\{R_n(\alpha, \hat{G}_n, F_n)\}_{n=1}^{\infty}$ $F_n$-stochastically equicontinuous at $\alpha_0$ if for each $\eta > 0$ and
$\varepsilon > 0$ there exists a neighborhood $U$ of $\alpha_0$ for which

$$\limsup_{n \to \infty} \mathbb{P}_{F_n} \left\{ \sup_{U} \left| R_n(\alpha, \hat{G}_n, F_n) - R_n(\alpha_0, \hat{G}_n, F_n) \right| > \eta \right\} < \varepsilon.$$ 

As the name suggests, with very high probability, an $F_n$-stochastically equicontinuous process will not vary wildly in a small neighborhood of a point. A very important consequence of that property is that if $\{\alpha_n(\hat{G}_n)\}$ is a sequence of random variables which converges in $F_n$-probability to $\alpha_0$, then

$$R_n(\alpha_n(\hat{G}_n), \hat{G}_n, F_n) - R_n(\alpha_0, \hat{G}_n, F_n) \to 0 \quad \text{in } F_n\text{-probability.}$$

Consequently, if $R_n(\alpha_0, \hat{G}_n, F_n)$ converges in distribution, then so does $R_n(\alpha_n(\hat{G}_n), \hat{G}_n, F_n)$ and their asymptotic distributions are the same. Thus, in the case of the trimmed mean process $M_n(\alpha, \hat{F}_n, F)$, we know by corollary 2.1 that $M_n(A_0(F), \hat{F}_n, F)$ converges in distribution to a normal distribution with variance $\sigma^2(A_0(F), F)$, and so to prove theorems 3.1, 3.3 and 3.4, we need to show that the trimmed mean process $M_n(\alpha, \hat{F}_n, F)$ is $F$-stochastically equicontinuous at $A_0(F)$, and prove lemmas 3.1 through 3.3.

When a bootstrap confidence interval is constructed, the third argument of $R_n$ is no longer fixed, but is random. Instead of sampling from the fixed distribution function $F$, one samples from a random distribution function, say $\hat{F}_n$. Most of the time $\hat{F}_n$ is the empirical distribution function $\hat{F}_n$. But our approach will be such that it will also be easy to consider other estimates of $F$ such as $\hat{F}_{n,\alpha_n}$ and $\hat{F}_n$ introduced in section 3.5.

The randomness of the third argument of $R_n$ certainly complicates the asymptotic theory. But Beran (1984) introduced a method that greatly simplifies the theory. In this context, it consists of showing that $R_n(\alpha_n(\hat{G}_n), \hat{G}_n, F_n)$ has the same asymptotic distribution as $R_n(\alpha_0, \hat{F}_n, F)$ for all sequences of fixed distribution functions $\{F_n\}$ which converge to $F$ in a certain sense. Then one must simply show that $\{\hat{F}_n\}$, or another random sequence which has the same distribution on an appropriate space, converges to $F$ in the aforementioned sense with probability one. More on that later. But for now, we shall therefore consider fixed sequences of distribution functions $\{F_n\}$.
To demonstrate that the bootstrap distribution \( R_n(\alpha_n(\hat{G}_n), \hat{G}_n, F_n) \) has the same asymptotic distribution as \( R_n(\alpha_0, \hat{F}_n, F) \), the first two steps consist of showing that \( R_n(\alpha_n(\hat{G}_n), \hat{G}_n, F_n) \) is \( F_n \)-stochastically equicontinuous and that \( \alpha_n(\hat{G}_n) \) converges in \( F_n \)-probability to \( \alpha_0 \). This implies that \( R_n(\alpha_n(\hat{G}_n), \hat{G}_n, F_n) - R_n(\alpha_0, \hat{G}_n, F_n) \) converges to 0 in \( F_n \)-probability as \( n \to \infty \). This, in turn, implies that \( R_n(\alpha_n(\hat{G}_n), \hat{G}_n, F_n) \) has the same asymptotic distribution as \( R_n(\alpha_0, \hat{G}_n, F_n) \) if the latter has a weak limit. The hope of the bootstrap method is that \( R_n(\alpha_0, \hat{G}_n, F_n) \) behaves asymptotically like \( R_n(\alpha_0, \hat{F}_n, F) \). So we must also show that the asymptotic distributions of these two random variables are identical. Thus three results need to be established before letting the fixed sequences \( \{F_n\} \) be random again. Let's formalize this into the following theorem.

**THEOREM 4.1:** Suppose that

1. \( \alpha_n(\hat{G}_n) \to \alpha_0 \) in \( F_n \)-probability,

2. \( R_n(\alpha, \hat{G}_n, F_n) \) is \( F_n \)-stochastically equicontinuous at \( \alpha_0 \), and,

3. The laws of \( R_n(\alpha_0, \hat{G}_n, F_n) \) and \( R_n(\alpha_0, \hat{F}_n, F) \) both converge weakly to the law of \( R(\alpha_0, F) \).

Then the law of \( R_n(\alpha_n(\hat{G}_n), \hat{G}_n, F_n) \) converges weakly to the law of \( R(\alpha_0, F) \).

Note that each of those three conditions further restrict the possible candidates \( \{F_n\} \). For instance, the first two conditions can be satisfied without the third one holding. Let \( F_{2n+1} \) and \( F_{2n} \) be the distribution functions of a normal random variable with variance 1 and variance 2, respectively. Then consider any \( \alpha_n(\hat{G}_n) \) such that the first condition is satisfied (e.g., the two random choices of chapter 3 would work since they are invariant under scale changes). As we will soon discover, the second condition is satisfied for \( M_n(\alpha, \hat{F}_n, F) \) when \( F \) is any normal distribution function. So, even though \( F_n \) alternates between two different distribution functions, \( M_n(\alpha, \hat{G}_n, F_n) \) will still satisfy the second condition. But clearly, \( M_n(\alpha_0, \hat{G}_n, F_n) \) does not have a weak limit since \( M_{2n}(\alpha_0, \hat{G}_{2n}, F_{2n}) \) and \( M_{2n+1}(\alpha_0, \hat{G}_{2n+1}, F_{2n+1}) \) both have weak limits which are not identical.
The third condition basically requires that the sequence of distribution functions \( \{F_n\} \) converge to \( F \) in some sense, whereas the second condition guarantees that the process cannot vary too much in a small neighborhood of \( \alpha_0 \). Therefore these two conditions combined imply that as long as the sampling distribution is close enough to \( F \) and that the selected value of \( \alpha \) is close enough to \( \alpha_0 \), the distribution function of the process evaluated at the approximate and exact values will not be very different.

Therefore the idea consists of defining a class \( CF \) of sequences of distribution functions \( \{F_n\} \) such that each \( \{F_n\} \in CF \) satisfies the three previously stated conditions. Since the first condition depends on the particular method of adaptation, so could the class \( CF \). So, by definition, each \( \{F_n\} \in CF \) satisfies the property that \( R_n(\alpha_n, \hat{G}_n, \hat{G}_n, F_n) \) has the same asymptotic distribution as \( R_n(\alpha_0, \hat{F}_n, F) \). Assuming that the asymptotic distribution is continuous (which is the case whenever the weak limit is normal), this result can be rephrased as follows: the supremum (over \( x \)) of the absolute difference between the distribution functions of \( R_n(\alpha_n, \hat{G}_n, \hat{G}_n, F_n) \) and its weak limit (evaluated at \( x \)) converges to 0. In symbols, let \( J_n(x, F_n) \) be the distribution function of \( R_n(\alpha_n, \hat{G}_n, \hat{G}_n, F_n) \) evaluated at \( x \), and \( J(x, F) \) be the distribution function of the asymptotic distribution of \( R_n(\alpha_0, \hat{F}_n, F) \) evaluated at \( x \). Then, for each \( \{F_n\} \in CF \),

\[
\sup_x |J_n(x, F_n) - J(x, F)| \to 0. \tag{4.1}
\]

But what about the case where \( \{\hat{F}_n\} \) is a random sequence of distribution functions, as in the bootstrap? The idea should now be clear: show that \( \{\hat{F}_n\} \in CF \) with probability 1. In such a case, the corresponding statement in (4.1) would be valid with probability 1. Unfortunately, often \( \{\hat{F}_n\} \) is not in \( CF \) with probability 1. But all is not lost as, in most cases, it is possible to define a random sequence of distribution functions \( \{\hat{F}_n^*\} \) on an appropriate probability space, using Skorohod's representation theorem, such that \( \{\hat{F}_n^*\} \in CF \) with probability 1, and \( \hat{F}_n^* \) and \( \hat{F}_n \) have the same distribution when regarded as random variables on some appropriate space (say \( D(-\infty, \infty) \)). Hence, (4.1) becomes

\[
\sup_x |J_n(x, \hat{F}_n^*) - J(x, F)| \to 0, \quad \text{with probability 1.}
\]
But since $\tilde{F}_n^*$ and $\tilde{F}_n$ have the same distribution, this implies that

$$\sup_x |J_n(x, \tilde{F}_n) - J(x, F)| \to 0,$$

in probability. \hfill (4.2)

This is a weaker type of convergence, but it is sufficient to guarantee that confidence intervals based on the bootstrap adaptive root $R_n(\alpha_n(\hat{G}_n), \hat{G}_n, \tilde{F}_n)$ asymptotically have the claimed coverage probability. For such a result, see theorem 1 of Beran (1984), or the proof of corollary 3.3 in the next section.

Beran's idea of working with a class of fixed distribution functions is very useful as it clearly outlines the properties that must be satisfied by estimates $\tilde{F}_n$ of $F$. Now let's concentrate on the adaptive choices that were introduced in chapter 3.

All three adaptive choices share the same structure. They all compute an estimate of a certain quantity for all values of $\alpha$ such that $n\alpha$ is an integer and $\alpha$ is in an interval. Their value is then the $\alpha$ which minimizes these estimates. In the case of Jaeckel's estimate, the quantity to be estimated is the asymptotic variance, whereas in the bootstrap estimates, it is the variance and the interquartile range of the finite sample distribution. Since it is assumed that the law of $R_n(\alpha, \hat{G}_n, F_n)$ converges weakly for each fixed value of $\alpha$ (to a continuous, strictly increasing distribution function), then the interquartile range of the finite sample distribution also converges to the interquartile range of the asymptotic distribution. Therefore, an estimate of the interquartile range of the finite sample distribution can be viewed as an estimate of the interquartile range of the asymptotic distribution, as it estimates a quantity which itself converges to the asymptotic interquartile range. Likewise, if it is known that the finite sample variance converges to the asymptotic variance, then an estimate of the finite sample variance can be viewed as an estimate of the asymptotic variance. Of course, this is an asymptotic view. In finite samples, one might do better by minimizing an estimate of the finite sample variance rather than an estimate of the asymptotic variance. Nevertheless, the two approaches would be asymptotically equivalent in such a case. So, in this section, we will view $S_{\text{Jae}}$, $S_{\text{Var}}^n$ and $S_{\text{Inq}}^n$ as estimates of an asymptotic quantity.
In general, consider estimates $S_n(\alpha, \hat{G}_n)$ (we are assuming for now that we have a sample of size $n$ from a fixed $F_n$). They are estimating a characteristic $S(\alpha, F)$ of the asymptotic distribution of $R_n(\alpha, \hat{F}_n, F)$, usually a characteristic of spread such as the variance or the interquartile range. Then, the adaptive choice is defined to be

$$\alpha_n(F_n) = \arg\min_{\alpha \in A_n} S_n(\alpha, \hat{G}_n),$$

where $A_n$ is a set that may depend on $n$ and usually is of finite cardinality, so as to make the method numerically feasible. In chapter 3, for instance, $A_n$ consists of all $\alpha$'s in a given interval such that $n\alpha$ is an integer.

The goal is to do as well as possible asymptotically. So in this case, we are interested in minimizing $S(\alpha, F)$ over a set of indices $A$. Therefore, let

$$\alpha_0(F) = \arg\min_{\alpha \in A} S(\alpha, F),$$

where $A$ need not be of finite cardinality. As was discussed before, it is therefore hoped that $\alpha_n(F_n)$ converges in $F_n$-probability to $\alpha_0(F)$. By definition of $S_n$, that result will depend on the characteristic being estimated, the smoothness of the process $R_n$, and the closeness of $F_n$ to $F$. The last two conditions have already been mentioned previously and are needed for the weak convergence of the adaptive estimator. Nevertheless, the key step in proving that result (and therefore results such as lemmas 3.1 through 3.3) is to show that

$$\sup_{\alpha \in A_n} |S_n(\alpha, \hat{G}_n) - S(\alpha, F)| \to 0, \quad \text{in } F_n\text{-probability.} \quad (4.3)$$

In other words, the estimates must converge to the estimands in probability, uniformly over all $\alpha$'s in $A$. If this result holds, then the smoothness of the argmin function will imply the desired result. As must be expected, the proof of a result like (4.3) depends to a large extent on the characteristic being estimated.

Let's proceed with the proofs of the results in chapter 3.
4.2. Proofs

In this section we will prove the results of chapter 3. The outline presented in the preceding section will be followed. Let's first describe the condition which will be required by the sequences \( \{F_n\} \) to be members of \( CF \).

**CONDITION (C):** Consider a sequence of fixed distribution functions \( \{F_n\} \), and a fixed distribution function \( F \). There exists a constant \( D \) such that for all large \( n \),

\[
 n^{1/2} \sup_x |F_n(x) - F(x)| \leq D.
\]

In the last section, it was argued that \( F_n \) must converge to \( F \) in some sense. Condition (C) defines the type of convergence that must be satisfied by \( F_n \). Now notice that condition (C), in the presence of assumption (A.2) implies the following condition (\( C^* \)):

\( (C^*) \) For all large \( n \),

\[
 n^{1/2} \sup_{[\alpha_0, 1-\alpha_0]} |F_n^{-1}(\alpha) - F^{-1}(\alpha)| \leq \frac{D + 1}{f_0},
\]

where \( f_0 \) is defined in (A.2).

Formally we have the following lemma.

**Lemma 4.1:** If \( n^{1/2} \sup_x |F_n(x) - F(x)| \leq D \) and \( F \) has a density \( f \) which is bounded below by \( f_0 > 0 \) on \( \{ x : \alpha_0 - \epsilon \leq F(x) \leq 1 - \alpha_0 + \epsilon \} \), for some \( \alpha_0 \geq 0 \) and some \( \epsilon > 0 \), then

\[
 n^{1/2} \sup_{[\alpha_0, 1-\alpha_0]} |F_n^{-1}(\alpha) - F^{-1}(\alpha)| \leq \frac{D + 1}{f_0}.
\]

**Proof:** Let \( \epsilon_n = \left( \frac{D + 1}{f_0} \right) n^{-1/2} \). Then note that by (C), for any \( \alpha \in [\alpha_0, 1 - \alpha_0] \)

\[
 F_n(F^{-1}(\alpha) - \epsilon_n) = F(F^{-1}(\alpha) - \epsilon_n) + \delta_n, \quad \text{where } |\delta_n| \leq Dn^{-1/2}
\]

\[
 = \alpha - \epsilon_n f(x^*) + \delta_n, \quad \text{where } x^* \in [F^{-1}(\alpha) - \epsilon_n, F^{-1}(\alpha)]
\]

\[
 \leq \alpha - f_0 \epsilon_n + |\delta_n|
\]

\[
 \leq \alpha - \frac{f_0(D + 1)n^{-1/2}}{f_0} + Dn^{-1/2}
\]

\[
 = \alpha - n^{-1/2}.
\]
Likewise, \( F_n(F^{-1}(\alpha) + \epsilon_n) \geq \alpha + n^{-1/2} \). Note that \( F(x) < p \) implies \( F^{-1}(p) \geq x \), and \( F(x) \geq p \) if and only if \( F^{-1}(p) \leq x \). Consequently, for all large \( n \),
\[
F_n(F^{-1}(\alpha) - \epsilon_n) < \alpha < F_n(F^{-1}(\alpha) + \epsilon_n)
\]
implies
\[
F^{-1}(\alpha) - \epsilon_n \leq F_n^{-1}(\alpha) \leq F^{-1}(\alpha) + \epsilon_n.
\]
i.e.,
\[
n^{1/2} \sup_{[\alpha_0,1-\alpha_0]} |F_n^{-1}(x) - F^{-1}(x)| \leq \frac{D+1}{f_0}.
\]

We are now in a position to define the class of distribution functions \( C_F \).

**Definition 4.2:** For distribution functions \( F \) that satisfy assumption A of chapter 3, define \( C_F \) to be the class of sequences of distribution functions \( \{F_n\} \) such that condition (C) is satisfied.

The next order of business consists of showing that the trimmed mean process \( M_n(\alpha, \hat{G}_n, F_n) \) is \( F_n \)-stochastically equicontinuous for any member \( \{F_n\} \) of \( C_F \). Formally, we have the following theorem.

**Theorem 4.2:** Let \( F \) satisfy assumption A, and suppose that \( \{F_n\} \) satisfies condition (C). Then \( M_n(\alpha, \hat{G}_n, F_n) \) is \( F_n \)-stochastically equicontinuous for any \( \alpha \in [\alpha_0, 1/2) \).

**Proof:** Let \( \alpha_1 \in [\alpha_0, 1/2) \) be given. Let also \( \epsilon, \eta > 0 \) be given. Consider any \( \alpha_1 < \alpha_2 < 1/2 \). Then
\[
|M_n(\alpha_2, \hat{G}_n, F_n) - M_n(\alpha_1, \hat{G}_n, F_n)|
\]
\[
= n^{1/2} \left| T(\alpha_2, \hat{G}_n) - T(\alpha_2, F_n) \right| - n^{1/2} \left| T(\alpha_1, \hat{G}_n) - T(\alpha_1, F_n) \right|
\]
\[
= n^{1/2} \left| T(\alpha_2, \hat{G}_n) - T(\alpha_1, \hat{G}_n) \right| - n^{1/2} \left| T(\alpha_2, F_n) - T(\alpha_1, F_n) \right|
\]
\begin{align*}
&= n^{1/2} \int_0^1 \hat{G}^{-1}_n(t) \left[ \frac{I(\alpha_2 \leq t \leq 1 - \alpha_2)}{1 - 2\alpha_2} - \frac{I(\alpha_1 \leq t \leq 1 - \alpha_1)}{1 - 2\alpha_1} \right] dt \\
&- n^{1/2} \int_0^1 F^{-1}_n(t) \left[ \frac{I(\alpha_2 \leq t \leq 1 - \alpha_2)}{1 - 2\alpha_2} - \frac{I(\alpha_1 \leq t \leq 1 - \alpha_1)}{1 - 2\alpha_1} \right] dt \\
&\leq n^{1/2} \int_{\alpha_1}^{1-\alpha_1} \left| \hat{G}^{-1}_n(t) - F^{-1}_n(t) \right| \left[ \frac{I(\alpha_2 \leq t \leq 1 - \alpha_2)}{1 - 2\alpha_2} - \frac{I(\alpha_1 \leq t \leq 1 - 2\alpha_1)}{1 - 2\alpha_1} \right] dt \\
&\leq n^{1/2} \sup_{t \in [\alpha, 1-\alpha]} \left| \hat{G}^{-1}_n(t) - F^{-1}_n(t) \right| \times \int_{\alpha_1}^{1-\alpha_1} \left[ \frac{I(\alpha_2 \leq t \leq 1 - \alpha_2)}{1 - 2\alpha_2} - \frac{I(\alpha_1 \leq t \leq 1 - 2\alpha_1)}{1 - 2\alpha_1} \right] dt. \quad (4.4)
\end{align*}

In order to bound (4.4), we need an upper bound for \( n^{1/2} \sup_{t \in [\alpha, 1-\alpha]} |\hat{G}^{-1}_n(t) - F^{-1}_n(t)| \). By Dvoretzky-Kiefer-Wolfowitz's inequality, (Dvoretzky, Kiefer, and Wolfowitz (1956)), \( \sup_x |\hat{G}_n(x) - F_n(x)| = O_p(n^{-1/2}) \). Thus there exists \( M < \infty \) such that for all large \( n \),

\begin{equation}
\text{Prob}_{F_n} \left\{ n^{1/2} \sup_x |\hat{G}_n(x) - F_n(x)| \leq M \right\} > 1 - \epsilon. \quad (4.5)
\end{equation}

Likewise, by condition (C), for all large \( n \)

\begin{equation}
n^{1/2} \sup_x |F_n(x) - F(x)| \leq D. \quad (4.6)
\end{equation}

Also, by condition (C*),

\begin{equation}
n^{1/2} \sup_{\alpha \in [\alpha_0, 1-\alpha_0]} |F^{-1}_n(\alpha) - F^{-1}(\alpha)| \leq \frac{D + 1}{f_0}, \quad (4.7)
\end{equation}

for all large \( n \). Using (4.6) and (4.7) along with an argument almost identical to the proof that (C) implies (C*), it is possible to show that (4.5) implies the following result:

\begin{equation}
\text{Prob}_{F_n} \left\{ n^{1/2} \sup_{t \in [\alpha_0, 1-\alpha_0]} |\hat{G}^{-1}_n(t) - F^{-1}_n(t)| \leq E \right\} > 1 - \epsilon, \quad (4.8)
\end{equation}

for all large \( n \) and a fixed constant \( E \).

Next we see that

\begin{align*}
\int_{\alpha_1}^{1-\alpha_1} \left| \frac{I(\alpha_2 \leq t \leq 1 - \alpha_2)}{1 - 2\alpha_2} - \frac{I(\alpha_1 \leq t \leq 1 - 2\alpha_1)}{1 - 2\alpha_1} \right| dt \\
= \int_{\alpha_1}^{\alpha_2} \frac{1}{1 - 2\alpha_1} dt + \int_{\alpha_2}^{1-2\alpha_2} \frac{1}{1 - 2\alpha_2} dt - \frac{1}{1 - 2\alpha_1} dt + \int_{1-\alpha_2}^{1-\alpha_1} \frac{1}{1 - 2\alpha_1} dt
\end{align*}
\[ = \frac{4(\alpha_2 - \alpha_1)}{1 - 2\alpha_1}. \]  

(4.9)

Letting \( \delta = \min((1/2)(1/2 - \alpha_1), (\eta/2)^{1-2\alpha_1}) \), and \( U = [\alpha_1 - \delta, \alpha_1 + \delta] \), then (4.6) and (4.9) imply

\[
\text{Prob}_F \left\{ \sup_{\alpha \in U} |M_n(\alpha, \hat{G}_n, F_n) - M_n(\alpha_1, \hat{G}_n, F_n)| < \eta \right\} > 1 - \epsilon,
\]

for all large \( n \), i.e., \( M_n(\alpha, \hat{G}_n, F_n) \) is \( F_n \)-stochastically equicontinuous at \( \alpha_1 \). \( \blacksquare \)

As we have seen, theorem 4.2 is one of the three results required to prove theorems 3.1 through 3.6. Let’s now turn on to the bootstrap adaptive trimmed mean based on a bootstrap estimate of the interquartile range. As it was indicated earlier in this chapter, the key element to show that \( \alpha_n^{\text{IQR}}(\hat{G}_n) \) converges in \( F_n \)-probability to \( A_0(F) \) is to demonstrate that \( \sup_{\alpha \in I_n} |S_n^{\text{IQR}}(\alpha, \hat{G}_n) - S^{\text{IQR}}(\alpha, F)| \to 0 \) in \( F_n \)-probability, where \( S^{\text{IQR}}(\alpha, F) \) is the interquartile range of the asymptotic distribution of the \( \alpha \)-trimmed mean of a sample from \( F \). But since \( S_n^{\text{IQR}}(\alpha, \hat{G}_n) \) is an interquartile range, and therefore the difference of two quantiles, such a result will be implied by the uniform weak convergence of the distribution function of \( M_n(\alpha, \hat{G}_n, F_n) \) to the distribution function of a normal with asymptotic variance \( \sigma^2(\alpha, F) \). It is interesting to note that an immediate corollary to this result is that \( M_n(A_0(F), \hat{G}_n, F_n) \) and \( M_n(A_0(F), \hat{E}_n, F) \) both converge to the same weak limit. Thus the next lemma is the key to establishing condition 1 for \( \alpha_n^{\text{IQR}} \) and also implies condition 2.

Recall from chapter 2 that \( L_n(x; \alpha, F_n) \) is the distribution function of \( M_n(\alpha, \hat{G}_n, F_n) \), i.e.,

\[
L_n(x; \alpha, F_n) = \text{Prob}_F \left\{ M_n(\alpha, \hat{G}_n, F_n) \leq x \right\},
\]

and that \( L(x; \alpha, F) = \Phi(x/\sigma(\alpha, F)) \) is its asymptotic distribution function.

**Lemma 4.2:** Let \( F \) satisfy assumptions (A.1) and (A.2), and suppose that \( \{F_n\} \) satisfies condition (C). If \( b_n \to 0 \), let’s further assume that \( F \) satisfies assumption (A.4).
Then if \( n^{1/2}b_n \to \infty \) as \( n \to \infty \),

\[
\sup_x \sup_{\alpha \in I_n} |L_n(x; \alpha, F_n) - L(x; \alpha, F)| \to 0, \quad \text{as } n \to \infty,
\]

where \( I_n = [\alpha_0, 1/2 - b_n] \).

**Sketch of the proof:** The idea of the proof is simple. Differentiate the process \( M_n \) for each value of \( \alpha \). Show that the linear approximation converges weakly uniformly in \( \alpha \). This involves showing that the asymptotic variance of the \( \alpha \)-trimmed mean based on \( F_n \) converges to \( \sigma^2(\alpha, F) \) uniformly in \( \alpha \). Then show that the remainder converges to 0 in \( F_n \)-probability uniformly in \( \alpha \).

**Proof:** The differentiable approach of Boos (1979) and Serfling (1980) will be used. Serfling uses this differentiable approach to show the weak convergence of a class of L-estimators that can be written in a functional form as \( U(F) = \int_0^1 F^{-1}(t)w(t) \, dt \), where \( w(t) \) is a weight function. Note that the \( \alpha \)-trimmed mean can be written as \( T(\alpha, F) = \int_0^1 F^{-1}(t)w_\alpha(t) \, dt \) for \( w_\alpha(t) = (1 - 2\alpha)^{-1}I(\alpha \leq t \leq 1 - \alpha) \) where \( I \) is the indicator function.

Now let \( Y_{n1}, \ldots, Y_{nn} \) be a sample of size \( n \) from \( F_n \) with empirical distribution function \( \hat{G}_n \). Then from Boos (1979), we can write

\[
M_n(\alpha, \hat{G}_n, F_n) = n^{1/2}dM_n(\alpha, \hat{G}_n, F_n) + n^{1/2}\Delta_n(\alpha, \hat{G}_n, F_n),
\]

where

\[
dM_n(\alpha, \hat{G}_n, F_n) = 1/n \sum_{1}^{n} h_\alpha(Y_{ni}; F_n),
\]

and

\[
h_\alpha(x; F_n) = -\int_{-\infty}^{\infty} \left[ I(y \geq x) - F_n(y) \right] w_\alpha(F_n(y)) \, dy
\]

\[
= -(1 - 2\alpha)^{-1} \int_{F_n^{-1}(1 - \alpha)}^{F_n^{-1}(\alpha)} \left[ I(y \geq x) - F_n(y) \right] \, dy.
\]

The remainder term \( n^{1/2}\Delta_n(\alpha, \hat{G}_n, F_n) \) is simply the difference between \( M_n(\alpha, \hat{G}_n, F_n) \) and the linear term \( n^{1/2}dM_n(\alpha, \hat{G}_n, F_n) \).
Boos (1979) has showed that whenever $F_n^{-1}$ is continuous at $\alpha$ and $1 - \alpha$, then $dM_n$ is the derivative of the $\alpha$-trimmed mean with respect to the sup-norm. But such an assumption about $F_n$ is not made. Therefore $dM_n$ need not be the derivative. Nevertheless, the decomposition (4.10) will still be very useful because, by condition $(C^*)$, $F_n^{-1}(\alpha)$ and $F_n^{-1}(1 - \alpha)$ converge respectively to $F^{-1}(\alpha)$ and $F^{-1}(1 - \alpha)$, and by assumption (A.2), $F^{-1}$ is continuous at both $\alpha$ and $1 - \alpha$, for $\alpha \in [\alpha_0, 1 - \alpha_0]$.

Let's now study the basic properties of the random variable $h_\alpha(Y; F_n)$. First, it is bounded. By (4.10), we see that

$$|h_\alpha(Y; F_n)| \leq 2 \left[ \frac{F_n^{-1}(1 - \alpha) - F_n^{-1}(\alpha)}{1 - 2\alpha} \right]$$

$$\leq 2 \left[ \frac{(D + 1)n^{-1/2}}{f_0(1 - 2\alpha)} + \frac{F_n^{-1}(1 - \alpha) - F_n^{-1}(\alpha)}{1 - 2\alpha} \right],$$

(4.12)

for all large $n$, since by condition $(C^*)$ $|F_n^{-1}(\alpha) - F_n^{-1}(\alpha)| \leq \left( \frac{D + 1}{f_0} \right)n^{-1/2}$.

Consider the function $(F_n^{-1}(1 - \alpha) - F_n^{-1}(\alpha)) / (1 - 2\alpha)$ for $\alpha \in [\alpha_0, 1/2)$. Since $\alpha_0 > 0$, this function could only be unbounded for $\alpha$ close to $1/2$. But it is easy to show that

$$\lim_{\alpha \to 1/2} (F_n^{-1}(1 - \alpha) - F_n^{-1}(\alpha)) / (1 - 2\alpha) = 1/f(0) \leq 1/f_0$$

by virtue of assumption (A.2). So let

$$E = \sup_{\alpha \in [\alpha_0, 1/2]} \left\{ \frac{F_n^{-1}(1 - \alpha) - F_n^{-1}(\alpha)}{1 - 2\alpha} \right\} < \infty.$$  

(4.13)

Then for all large $n$, $n$

$$\sup_{\alpha \in I_n} |h_\alpha(Y; F_n)| \leq 2 \left[ \frac{D + 1}{f_0b_n} n^{-1/2} + E \right]$$

$$\leq \text{constant}.$$  

(4.14)

Second note that $E_h \cdot h_\alpha(Y; F_n) = 0$ since

$$E_h \cdot h_\alpha(Y; F_n) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [I(y \geq x) - F_n(y)] w_\alpha(F_n(y)) \, dy \, dF_n(x)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [I(y \geq x) - F_n(y)] dF_n(x) w_\alpha(F_n(y)) \, dy$$

$$= 0 \quad \text{by Fubini's Theorem.}$$
Third, let \( v(\alpha, F_n) \) be the finite variance of \( h_\alpha(Y; F_n) \). Then using Fubini’s Theorem in a similar fashion yields the following formula:

\[
v(\alpha, F_n) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F_n(\min(s, t)) - F_n(s) F_n(t)] w_\alpha(F_n(s)) w_\alpha(F_n(t)) \, ds \, dt. \tag{4.15}
\]

The next step consists of showing that the linear term \( n^{1/2} dM_n(\alpha, \hat{G}_n, F_n) \) converges weakly, uniformly in \( \alpha \). First, \( k^{-1/2} v^{-1/2}(\alpha, F_n) \sum_1^k h_\alpha(Y_i; F_n) \) converges weakly to the standard normal distribution as \( k \to \infty \) (here \( F_n \) is fixed) by the central limit theorem. But it is the corresponding result with \( \sigma(\alpha, F) \) replacing \( v^{1/2}(\alpha, F_n) \) which is desired. Obviously, the convergence of \( v^{1/2}(\alpha, F_n) \) to \( \sigma(\alpha, F) \) would imply such a result. It will turn out that if this convergence is uniform over \( \alpha \) and that \( \sigma^2(\alpha, F) \) is bounded away from 0 as a function of \( \alpha \), then the weak convergence will also be uniform over \( \alpha \) and \( x \) by the Berry-Esseen Theorem. Remember that from (4.14) \( h_\alpha(Y; F_n) \) is bounded above uniformly in \( \alpha \) and, therefore, so is its third moment. Thus let’s show that \( v(\alpha, F_n) \to \sigma^2(\alpha, F) \) uniformly in \( \alpha \).

First, note that under assumption (A.2) \( F \) has a continuous inverse at \( \alpha \) and \( 1 - \alpha \), and so the differential approach of Boos is applicable to \( F \). Consequently \( v(\alpha, F) \), the finite sample variance of \( h_\alpha(Y; F) \), must also be the asymptotic variance of the \( \alpha \)-trimmed mean which is equal to \( \sigma^2(\alpha, F) \), as defined in (2.2). Thus by (4.15),

\[
|v(\alpha, F_n) - \sigma^2(\alpha, F)| = \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F_n(\min(s, t)) - F_n(s) F_n(t)] w_\alpha(F_n(s)) w_\alpha(F_n(t)) \, ds \, dt \\
- \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F(\min(s, t)) - F(s) F(t)] w_\alpha(F(s)) w_\alpha(F(t)) \, ds \, dt \right| \\
\leq \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ [F_n(\min(s, t)) - F_n(s) F_n(t)] - [F(\min(s, t)) - F(s) F(t)] \right\} \\
w_\alpha(F_n(s)) w_\alpha(F_n(t)) \, ds \, dt \right| \\
+ \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F(\min(s, t)) - F(s) F(t)] \left\{ w_\alpha(F_n(s)) w_\alpha(F_n(t)) \\
- w_\alpha(F(s)) w_\alpha(F(t)) \right\} \, ds \, dt \right|
\]
\[
\leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w_{\alpha}(F_n(s)) w_{\alpha}(F_n(t)) \left\{ |F_n(\min(s,t)) - F(\min(s,t))| + |F_n(s) - F(s)| \\
+ |F_n(t) - F(t)| \right\} ds \, dt \\
+ 2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left| w_{\alpha}(F_n(s)) w_{\alpha}(F_n(t)) - w_{\alpha}(F(s)) w_{\alpha}(F(t)) \right| ds \, dt \\
= T^1_n(\alpha) + T^2_n(\alpha).
\]

(4.16)

To bound \(T^1_n(\alpha)\) note that by condition (C), \(|F_n(s) - F(s)| \leq n^{-1/2}D\) for all large \(n\). Therefore,
\[
T^1_n(\alpha) \leq 3n^{-1/2}D \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w_{\alpha}(F_n(s)) w_{\alpha}(F_n(t)) \, ds \, dt \\
= 3n^{-1/2}D \left\{ \frac{F_n^{-1}(1 - \alpha) - F_n^{-1}(\alpha)}{1 - 2\alpha} \right\}^2,
\]

and so using the same trick as in (4.12),
\[
\sup_{\alpha \in \Lambda_n} T^1_n(\alpha) \leq 3n^{-1/2}D \sup_{\alpha \in \Lambda_n} \left\{ \frac{F_n^{-1}(1 - \alpha) - F_n^{-1}(\alpha)}{1 - 2\alpha} \right\}^2 \\
\leq 3n^{-1/2}D \left[ \frac{(D + 1)^2 n^{-1}}{f_0^2 b_n^2} + \frac{(D + 1)n^{-1/2}E}{f_0 b_n} + E^2 \right],
\]

(4.17)

where \(E\) is defined in (4.13). Now by assumption, \(n^{1/2}b_n \to \infty\) as \(n \to \infty\), and so the first term in (4.16) converges to 0 uniformly in \(\alpha\).

As for \(T^2_n(\alpha)\), note that the integrand is mostly equal to 0 since \(F_n(s)\) and \(F(s)\) are uniformly close and \(w_{\alpha}(\cdot)\) is a constant times the indicator of the interval \([\alpha, 1 - \alpha]\). More precisely, let
\[
S_n(\alpha) = \{ s : \alpha - \delta_n \leq F(s) \leq \alpha + \delta_n \text{ or } 1 - \alpha - \delta_n \leq F(s) \leq 1 - \alpha + \delta_n \},
\]

(4.18)

where \(\delta_n = \left( \frac{D+1}{f_0} \right) n^{-1/2}\). Then clearly, for every \(s \notin S_n(\alpha)\), \(w_{\alpha}(F_n(s)) = w_{\alpha}(F(s))\), since by condition (C), \(|F_n(s) - F(s)| \leq Dn^{1/2}\). Therefore,
\[
T^2_n(\alpha) \leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w_{\alpha}(F_n(s)) \left| w_{\alpha}(F_n(t)) - w_{\alpha}(F(t)) \right| \\
+ w_{\alpha}(F(t)) \left| w_{\alpha}(F_n(s)) - w_{\alpha}(F(s)) \right| ds \, dt \\
\leq 2 \left[ \frac{F_n^{-1}(1 - \alpha) - F_n^{-1}(\alpha)}{(1 - 2\alpha)^2} \right] \text{Leb}(S_n(\alpha))
\]

where \(\text{Leb}(S_n(\alpha))\) is the Lebesgue measure of \(S_n(\alpha)\).
\[ F^{-1}(1 - \alpha) - F^{-1}(\alpha) \]
\[ \frac{Leb(S_n(\alpha))}{(1 - 2\alpha)^2} \]
\[ \left( \frac{\delta_n}{f(\alpha^*)} - \frac{\delta_n}{f(\alpha^{**})} + \frac{\delta_n}{f(\alpha^{***})} - \frac{\delta_n}{f(\alpha^{****})} \right), \]
(4.19)

where \( Leb(S) \) is the Lebesgue measure of the set \( S \). Note that,

\[ Leb(S_n(\alpha)) = F^{-1}(\alpha + \delta_n) - F^{-1}(\alpha - \delta_n) + F^{-1}(1 - \alpha + \delta_n) - F^{-1}(1 - \alpha - \delta_n) \]
\[ = \frac{\delta_n}{f(\alpha^*)} - \frac{\delta_n}{f(\alpha^{**})} + \frac{\delta_n}{f(\alpha^{***})} - \frac{\delta_n}{f(\alpha^{****})}, \]

by the mean value theorem where \( \alpha^* \in (\alpha, \alpha + \delta_n), \alpha^{**} \in (\alpha - \delta_n, \alpha) \), and so on. But by assumption (A.2) the right hand side is bounded above by \( 4\delta_n/f_0 \), so

\[ Leb(S_n(\alpha)) \leq \frac{4\delta_n}{f_0}. \]
(4.20)

By condition (C*), \( |F^{-1}(\alpha) - F^{-1}(\alpha)| \leq \left( \frac{D+1}{f_0} \right) n^{-1/2} \), and so using (4.19), we get

\[ \sup_{\alpha \in I_n} \frac{F_n^2(\alpha)}{f_0(1 - 2\alpha)} \leq \frac{4\delta_n}{f_0} \left[ \frac{(D + 1)n^{-1/2}}{f_0(1 - 2\alpha)} + \frac{F^{-1}(1 - \alpha) - F^{-1}(\alpha)}{1 - 2\alpha} \right] \]
\[ \leq 4(D + 1)n^{-1/2} \left[ \frac{(D + 1)n^{-1/2}}{f_0b_n} + E \right], \]
(4.21)

where \( E \) is defined in (4.13). Hence, since by assumption \( n^{1/2}b_n \to \infty \) as \( n \to \infty \), the right hand side of (4.21) converges to 0.

Piecing (4.17) and (4.21) together yields the desired result, namely

\[ \sup_{\alpha \in I_n} |\nu(\alpha, F_n) - \sigma^2(\alpha, F)| \to 0 \quad \text{as } n \to \infty. \]
(4.22)

So the asymptotic variance of the \( \alpha \)-trimmed mean under \( F_n \) converges to the asymptotic variance under \( F \), uniformly in \( \alpha \).

It will now be shown that \( \sigma^2(\alpha, F) \) is uniformly bounded below away from 0. Let \( \alpha_0 < \alpha_1 < 1/2 \), then equation (2.2) implies that for any \( \alpha \in [\alpha_0, \alpha_1] \),

\[ \sigma^2(\alpha, F) > (1 - 2\alpha_1)^{-2} \left[ \int_{F^{-1}(\alpha_0)}^{F^{-1}(1 - \alpha_0)} (x - \theta)^2 dF(x) + 2\alpha_1(F^{-1}(\alpha_0) - \theta)^2 \right] \]
\[ > 0. \]
(4.23)

This means that for any \( \alpha_1 < 1/2, \sigma^2(\alpha, F) \) is uniformly bounded below away from 0 for all \( \alpha \in [\alpha_0, \alpha_1] \). So unless \( \lim_{\alpha \to 1/2} \sigma^2(\alpha, F) = 0 \), \( \sigma^2(\alpha, F) \) is uniformly bounded below.
away from 0 for all $\alpha \in [\alpha_0, 1/2)$. Using (2.2) and the mean value theorem, it is easy to show that $\lim_{n \to 1/2} \alpha^2(\alpha, F) = 1/(4f^2(0))$, i.e., it is equal to the asymptotic variance of the median, as it should be expected. But by assumption (A.4), $f(0)$ is finite, and so $\sigma^2(\alpha, F)$ is uniformly bounded below away from 0. Of course this result is also true if $b_n$ is equal to the same constant for all $n$, even if assumption (A.4) is not met, i.e.,

$$\inf_{\alpha_0 \leq \alpha \leq \alpha_1} \sigma^2(\alpha, F) > 0,$$

where $\alpha_1 = 1/2 - b_n$. In either case, the last two results imply that

$$\sup_{\alpha \in I_n} \left| \frac{\sigma^2(\alpha, F) - v(\alpha, F_n)}{\sigma^2(\alpha, F)v(\alpha, F_n)} \right| \to 0, \quad \text{as } n \to \infty. \quad (4.24)$$

Therefore,

$$\leq C n^{-1/2} + \sup_{\alpha \in I_n} \left| \frac{\sigma(\alpha, F) - v^{1/2}(\alpha, F_n)}{\sigma(\alpha, F)v^{1/2}(\alpha, F_n)} \right| \frac{|x| e^{-(1/2)(x^*)^2}}{\sqrt{2\pi}},$$

by the Berry-Esseen Theorem, where $x^*$ is between $x/\sigma(\alpha, F)$ and $x/v^{1/2}(\alpha, F_n)$.

By an argument similar to (4.23), it is clear that $\sigma(\alpha, F)$ is also bounded above over $[\alpha_0, 1/2)$, say by $M$. Hence for all large $n$, $x^* > x/(2M)$. Therefore,

$$\leq C n^{-1/2} + \sup_{\alpha \in I_n} \left| \frac{\sigma(\alpha, F) - v^{1/2}(\alpha, F_n)}{\sigma(\alpha, F)v^{1/2}(\alpha, F_n)} \right| \frac{|x| e^{-x^2/(8M)}}{\sqrt{2\pi}}.$$

But using (4.24) and the fact that $|x| e^{-x^2/(8M)}$ is uniformly bounded in $x$, the right hand side converges to 0. In words, this result says that $n^{1/2} dM_n(\alpha, G_n, F_n)$ converges weakly to the Gaussian distribution with variance $\sigma^2(\alpha, F)$, uniformly in $\alpha$. 
Next, it must be shown that the error term $n^{1/2} \Delta_n(\alpha, \hat{G}_n, F_n)$ converges to 0 in $F_n$-probability uniformly in $\alpha$. It turns out that

$$
\Delta_n(\alpha, \hat{G}_n, F_n) = - \int_{-\infty}^{\infty} \left[ W_\alpha(\hat{G}_n(x)) - W_\alpha(F_n(x)) - w_\alpha(F_n(x)) \left( \hat{G}_n(x) - F_n(x) \right) \right] dx
$$

where

$$
K_\alpha(x; G, F) = \begin{cases} 
\frac{W_\alpha(G(x)) - W_\alpha(F(x))}{G(x) - F(x)} - w_\alpha(F(x)), & \text{if } G(x) \neq F(x); \\
0, & \text{otherwise},
\end{cases}
$$

and

$$
W_\alpha(x) = \int_0^x w_\alpha(u) du.
$$

Hence, the following inequality is obviously satisfied

$$
|\Delta_n(\alpha, \hat{G}_n, F_n)| \leq ||K_\alpha(\cdot; \hat{G}_n, F_n)||_{L_1} ||\hat{G}_n - F_n||_\infty,
$$

(4.25)

where $||h||_\infty = \sup_x |h(x)|$ and $||h||_{L_1} = \int |h(x)| dx$. By Dvoretzky-Kiefer-Wolfowitz’s inequality, (DKW’s inequality), (Dvoretzky, Kiefer, and Wolfowitz (1956)), $||\hat{G}_n - F_n||_\infty = O_p(n^{-1/2})$. Therefore, to obtain

$$
\sup_{\alpha \in I_n} |\Delta_n(\alpha, \hat{G}_n, F_n)| = o_p(n^{-1/2}),
$$

it is sufficient to show that

$$
\sup_{\alpha \in I_n} ||K_\alpha(\cdot; \hat{G}_n, F_n)||_{L_1} \rightarrow 0 \quad \text{in } F_n\text{-probability}.
$$

In order to prove that result, let’s first notice that by the mean value theorem,

$$
K_\alpha(x; \hat{G}_n, F_n) = w_\alpha(y_n(x)) - w_\alpha(F_n(x)) \text{ for some } y_n(x) \text{ between } \hat{G}_n(x) \text{ and } F_n(x).
$$

By DKW’s inequality, $\text{Prob}_{F_n}(||\hat{G}_n - F_n||_\infty \leq M_1 n^{-1/2}) \geq 1 - C \exp\{-M_1^2\}$, for all $M_1$, all $n$, and a constant $C$. Also by condition (C), there exists $D$ such that for all $n$,

$$
||F_n - F||_\infty \leq Dn^{-1/2}.
$$

Hence, letting $M_2 = 2 \max(D, M_1)$ ($= 2M_1$, for $M_1 \geq D$), consider the set

$$
Q_n = \{ \hat{G}_n : \max(||\hat{G}_n - F||_\infty, ||\hat{G}_n - F_n||_\infty, ||F_n - F||_\infty) \leq M_2 n^{-1/2} \}.
$$
Then,

\[ Prob_{F_n}\{Q_n\} \geq 1 - C \exp\{-M_1^2\}. \]

So, letting \(\xi_n = M_2n^{-1/2}\), consider the set \(S_n(\alpha)\) defined in (4.18). It is clear that for \(\hat{G}_n \in Q_n\), if \(x \notin S_n(\alpha)\), then \(w_\alpha(y_n(x)) = w_\alpha(F_n(x))\) and so \(K_\alpha(x; \hat{G}_n, F_n) = 0\).

Thus, for all \(\hat{G}_n \in Q_n\),

\[ \sup_{\alpha \in I_n} \int_{-\infty}^{\infty} |K_\alpha(x; \hat{G}_n, F_n)| \, dx = \sup_{\alpha \in I_n} \int_{S_n(\alpha)} |K_\alpha(x; \hat{G}_n, F_n)| \, dx. \]

But for \(x \in S_n(\alpha), \sup_{\alpha \in I_n} |K_\alpha(x; \hat{G}_n, F_n)| \leq 2 \sup_{\alpha \in I_n} \|w_\alpha\|_\infty \leq 2b_n^{-1}.\) Thus for all \(n\) and all \(\hat{G}_n \in Q_n\),

\[ \sup_{\alpha \in I_n} \|K_\alpha(\cdot; \hat{G}_n, F_n)\|_{L_1} \leq 2b_n^{-1} \sup_{\alpha \in I_n} \text{Leb}(S_n(\alpha)) \]

\[ \leq 8b_n^{-1}n^{-1/2}M_2/f_0, \]

as shown in (4.20). But since \(Prob_{F_n}\{Q_n\} \geq 1 - C \exp\{-M_1^2\},\)

\[ Prob_{F_n}\{\sup_{\alpha \in I_n} \|K_\alpha(\cdot; \hat{G}_n, F_n)\|_{L_1} \leq 8b_n^{-1}n^{-1/2}M_2/f_0\} \geq 1 - C \exp\{-M_1^2\}, \]

i.e.,

\[ Prob_{F_n}\{\sup_{\alpha \in I_n} \|K_\alpha(\cdot; \hat{G}_n, F_n)\|_{L_1} \geq 8b_n^{-1}n^{-1/2}M_2/f_0\} \leq C \exp\{-M_1^2\}, \] (4.26)

In particular, since \(n^{1/2}b_n \to \infty, \sup_{\alpha \in I_n} \|K_\alpha(\cdot; \hat{G}_n, F_n)\|_{L_1} \to 0,\) in \(F_n\)-probability.

This implies that \(\sup_{\alpha \in I_n} |\Delta_n(\alpha, \hat{G}_n, F_n)| = o_p(n^{-1/2})\) which completes the proof.

An immediate consequence of this lemma is that \(M_n(\alpha, \hat{G}_n, F_n)\) and \(M_n(\alpha, \hat{F}_n, F)\) have the same asymptotic distribution, where \(\hat{F}_n\) is a sample of size \(n\) from \(F\).

**Corollary 4.1:** Let \(F\) satisfy assumptions (A.1) and (A.2), and suppose that \(\{F_n\}\) satisfies condition (C). If \(b_n \to 0\), let's further assume that \(F\) satisfies assumption (A.4). Then if \(n^{1/2}b_n \to \infty\) as \(n \to \infty\), the asymptotic distributions of \(M_n(\alpha, \hat{G}_n, F_n)\) and \(M_n(\alpha, \hat{F}_n, F)\) are identical.
Let us now make a few remarks.

**Remark 4.1:** It is important to notice that $w_\alpha(t)$, being a multiple of an indicator function, is not smooth at $\alpha$ and $1 - \alpha$. This fact is at the heart of the non-normal asymptotic distribution of the $\alpha$-trimmed mean whenever the quantile function is not continuous at $\alpha$ or $1 - \alpha$. Some authors have suggested using a smoother function. See Crow and Siddiqui (1967), and Stigler (1973). Stigler's proposed solution to this problem, for instance, is the following function:

$$w_\alpha(u) = \begin{cases} 
(u - \frac{\alpha}{2}) \frac{2h_\alpha}{\alpha}, & \text{if } \frac{\alpha}{2} \leq u \leq \alpha; \\
h_\alpha, & \text{if } \alpha \leq 1 - \alpha; \\
(1 - \frac{\alpha}{2} - u) \frac{2h_\alpha}{\alpha}, & \text{if } 1 - \alpha \leq u \leq 1 - \frac{\alpha}{2}; \\
0, & \text{otherwise},
\end{cases}$$

where $h_\alpha = 2(2 - 3\alpha)^{-1}$. A statistic based on that function would be asymptotically normal for any distribution. Moreover the arguments in the proofs of theorem 4.2 and lemma 4.2 would be valid with very little modifications even in the absence of assumption (A.2).

**Remark 4.2:** Throughout, it is assumed that the smallest trimming proportion under consideration, $\alpha_0$, is strictly positive. Hence the mean is not one of the possible candidates. This has allowed us to conveniently bound integrals like $\int_0^1 w_\alpha(F_n(t)) dt = (F_n^{-1}(1 - \alpha) - F_n^{-1}(\alpha))/(1 - 2\alpha)$ by $F_n^{-1}(1 - \alpha_0) - F_n^{-1}(\alpha_0)$ which converges to $F^{-1}(1 - \alpha_0) - F^{-1}(\alpha_0) < \infty$. Obviously, if $F$ is supported by a compact set, no such restriction on $\alpha_0$ would be required (although assumption (A.2) would still need to be satisfied, namely that the density be bounded below away from 0 on all of its support).

In general, it might be possible to let the smallest trimming proportion decrease to 0 just like the largest one tends to 1/2. The upper bound $F^{-1}(1 - \alpha_0) - F^{-1}(\alpha_0)$ is usually multiplied by a quantity, say $c_n$, which converges to 0 at a certain rate, in part determined by the rate at which the largest trimming proportion converges to 1/2. Therefore, it might be possible to let $\alpha_n$ tend to 0 at such a rate that $(F^{-1}(1 - \alpha_n) - F^{-1}(\alpha_n))c_n$
would still converge to 0. But that rate would then depend on the distribution function \( F \), a highly undesirable affair from a practical point of view. It is interesting to point out that the rate at which the smallest trimming proportion could converge to 0 would be in competition with the rate at which the largest trimming proportion would converge to 1/2.

In any case, extra conditions on \( F \) would be required. Remember from chapter 2 that for any \( \alpha > 0 \), the \( \alpha \)-trimmed mean is asymptotically normal for any distribution with continuous quantiles at \( \alpha \) and \( 1 - \alpha \). But we also know that the mean will be asymptotically normal only when the first two moments are finite. Looking at formula (2.2), it is clear that the asymptotic variance of the \( \alpha \)-trimmed mean diverges as \( \alpha \to 0 \) whenever the second moment does not exist. Hence it would be unlikely that one could get a result like lemma 4.2 with \( \alpha_0 \to 0 \) if the asymptotic variances are not uniformly bounded. Consequently, for all of these reasons, this approach is not pursued any further.

The next lemma uses lemma 4.2 to show that \( S_n^{\text{Iqr}}(\alpha, G_n) \) and \( S^{\text{Iqr}}(\alpha, F) \) are uniformly close whenever \( \{G_n\} \) satisfies condition (C).

**Lemma 4.3:** Let \( F \) satisfy assumptions (A.1) and (A.2), and suppose that \( \{G_n\} \) satisfies condition (C). Assume that \( n^{1/2}b_n \to \infty \) as \( n \to \infty \). If \( b_n \to 0 \), let's further assume that \( F \) satisfies assumption (A.4). Then

\[
\sup_{\alpha \in I_n} |S_n^{\text{Iqr}}(\alpha, G_n) - S^{\text{Iqr}}(\alpha, F)| \to 0, \quad \text{as } n \to \infty,
\]

where \( S^{\text{Iqr}}(\alpha, F) = L^{-1}(.75; \alpha, F) - L^{-1}(.25; \alpha, F) \), i.e., \( S^{\text{Iqr}}(\alpha, F) \) is the interquartile range of the asymptotic distribution of \( M_n(\alpha, \hat{\alpha}, F) \). Again \( I_n = [\alpha_0, 1/2 - b_n] \).

**Proof:** Let \( \epsilon > 0 \) be given. By lemma 4.2, there exists \( N_1 \) such that for all \( n \geq N_1 \),

\[
\sup_{x} \sup_{\alpha \in I_n} |L_n(x; \alpha, G_n) - L(x; \alpha, F)| < \epsilon.
\]

Therefore, for fixed \( 0 < y < 1 \), there exists \( N_2 \) such that for all \( n \geq N_2 \),

\[
L_n(L^{-1}(y; \alpha, F) - \epsilon; \alpha, G_n) < y < L_n(L^{-1}(y; \alpha, F) + \epsilon; \alpha, G_n), \quad \text{for all } \alpha \in I_n.
\]
Hence, as in the proof of lemma 4.1, for all \( n \geq N_2 \),
\[
L^{-1}(y; \alpha, F) - \epsilon \leq L_n^{-1}(y; \alpha, G_n) \leq L^{-1}(y; \alpha, F) + \epsilon, \quad \text{for all } \alpha \in I_n,
\]
i.e., for all \( n \geq N_2 \),
\[
\sup_{\alpha \in I_n} |L_n^{-1}(y; \alpha, G_n) - L^{-1}(y; \alpha, F)| \leq \epsilon, \quad (4.27)
\]
by the continuity of \( L \). But \( S_n^{\text{larg}}(\alpha, G_n) = L_n^{-1}(.75; \alpha, G_n) - L_n^{-1}(.25; \alpha, G_n) \), and so (4.27) implies that
\[
\sup_{\alpha \in I_n} |S_n^{\text{larg}}(\alpha, G_n) - S^{\text{larg}}(\alpha, F)| \to 0, \quad \text{as } n \to \infty.
\]

The next result is crucial. It says that if \( \{F_n\} \) satisfies condition (C) and that \( \hat{G}_n \) is the empirical distribution function of a sample of size \( n \) from \( F_n \), then there exists \( \hat{G}_n \) (on an appropriate probability space) with the same distribution as \( \hat{G}_n \) such that \( \{\hat{G}_n\} \) satisfies condition (C) almost surely. This result allows us to replace sequences of fixed distribution functions \( \{G_n\} \) by the sequence of (random) empirical distribution functions.

**Lemma 4.4:** Let \( \{F_n\} \) satisfy condition (C) for some distribution function \( F \). Let \( \hat{G}_n \) be the empirical distribution function of a sample of size \( n \) from \( F_n \). Then there exists \( \{\hat{G}_n\} \) with the same distribution as \( \{\hat{G}_n\} \) such that it satisfies condition (C) almost surely.

**Proof:** Let \( \{\xi_i\} \) be a sequence of independent random variables each having a uniform distribution on the interval \((0, 1)\) such that \( \|U_n - U\|_\infty \to 0 \) almost surely where \( \|x\|_\infty = \sup_r |x(r)| \), \( U_n(t) = n^{-1/2} \sum_{i=1}^n [I\{\xi_i \leq t\} - t] \) for all \( 0 \leq t \leq 1 \) and \( U \) is a Brownian bridge. See Shorack and Wellner (1986), Theorem 3.1.1 page 93 for the existence of such a process. This is a generalization of Skorohod's representation theorem.

Define the mappings \( H_n \) and \( H \) as follows:
\[
(H_n x)(r) = x(F_n(r)) \quad \text{and} \quad (H x)(r) = x(F(r)).
\]
Then,

\[ \|H_n U_n - H U\|_\infty = \sup_r |U_n(F_n(r)) - U(F(r))| \]

\[ \leq \sup_r |U_n(F_n(r)) - U(F_n(r))| + \sup_r |U(F_n(r)) - U(F(r))| \]

\[ \leq \|U_n - U\|_\infty + \sup_r |U(F_n(r)) - U(F(r))| \]

\[ \to 0, \quad \text{almost surely,} \]

since \( U \) is continuous on a compact set and \( \sup_r |F_n(r) - F(r)| \to 0 \) by condition (C).

Note that \( Y_{n,1}, \ldots, Y_{n,n} \), where \( Y_{n,i} = F_n^{-1}(\xi_i) \), is a sample of independently and identically distributed random variables with distribution \( F_n \). Let \( \tilde{G}_n(x) \) be the empirical distribution function of the \( Y_{n,i} \)’s. Note that \( \tilde{G}_n \) and \( \hat{G}_n \) have the same distribution. Then

\[ (H_n U_n)(r) = U_n(F_n(r)) \]

\[ = n^{-1/2} \sum_{i=1}^{n} [I\{\xi_i \leq F_n(r)\} - F_n(r)] \]

\[ = n^{-1/2} \sum_{i=1}^{n} [I\{F_n^{-1}(\xi_i) \leq r\} - F_n(r)] \]

\[ = n^{1/2}(\tilde{G}_n(r) - F_n(r)). \]

Therefore \( \|H_n U_n - H U\|_\infty \to 0 \) almost surely implies that

\[ \sup_r |n^{1/2}[\tilde{G}_n(r) - F_n(r)] - U(F(r))| \to 0, \quad \text{almost surely.} \quad (4.28) \]

But since the random paths of \( U \) are almost surely continuous on the compact set \([0,1]\), then \( U(F(r)) \) is almost surely finite. Therefore this fact and (4.28) imply that for almost every sample paths of \( \tilde{G}_n \), there exists \( D_1 < \infty \) (which depends on the sample path) such that for all large \( n \),

\[ n^{1/2} \sup_r |\tilde{G}_n(r) - F_n(r)| \leq D_1. \]

Moreover, by condition (C) there exists \( D_2 \) such that for all large \( n \),

\[ n^{1/2} \sup_r |F_n(r) - F(r)| \leq D_2. \]

Consequently, for almost every sample paths of \( \tilde{G}_n \), there exists \( D = D_1 + D_2 \) such that
for all large \( n \),
\[
n^{1/2} \sup_r |\tilde{G}_n(r) - F(r)| \leq D,
\]
i.e., \( \{\tilde{G}_n\} \) satisfies condition (C) almost surely.

We now return to lemma 4.3. Let \( \{F_n\} \in C_F \). Let \( \hat{G}_n \) be the empirical distribution function of a sample of size \( n \) from \( F_n \). The next lemma claims that \( S_{n}^{\text{lqr}}(\alpha, \hat{G}_n) \) and \( S_{n}^{\text{lqr}}(\alpha, F) \) are uniformly close in \( F_n \)-probability.

**Lemma 4.5:** Let \( F \) satisfy assumptions (A.1) and (A.2), and suppose that \( \{F_n\} \) satisfies condition (C). Assume that \( n^{1/2} b_n \to \infty \) as \( n \to \infty \). If \( b_n \to 0 \), let’s further assume that \( F \) satisfies assumption (A.4). Let \( \hat{G}_n \) be the empirical distribution function of a sample of size \( n \) from \( F_n \). Then
\[
\sup_{\alpha \in I_n} |S_{n}^{\text{lqr}}(\alpha, \hat{G}_n) - S_{n}^{\text{lqr}}(\alpha, F)| \to 0, \quad \text{in } F_n\text{-probability},
\]
where \( I_n = [a_0, 1/2 - b_n] \).

**Proof:** Let \( \{\tilde{G}_n\} \) be as defined in lemma 4.4. Then from lemma 4.3, it is possible to conclude that
\[
\sup_{\alpha \in I_n} |S_{n}^{\text{lqr}}(\alpha, \tilde{G}_n) - S_{n}^{\text{lqr}}(\alpha, F)| \to 0, \quad \text{almost surely}.
\]
But since \( \{\tilde{G}_n\} \) has the same distribution as \( \{\hat{G}_n\} \), we immediately conclude the following weaker statement:
\[
\sup_{\alpha \in I_n} |S_{n}^{\text{lqr}}(\alpha, \hat{G}_n) - S_{n}^{\text{lqr}}(\alpha, F)| \to 0, \quad \text{in } F_n\text{-probability}.
\]

Note that since \( \{F_n\} \) satisfies condition (C), then from lemma 4.3
\[
\sup_{\alpha \in I_n} |S_{n}^{\text{lqr}}(\alpha, F_n) - S_{n}^{\text{lqr}}(\alpha, F)| \to 0, \quad \text{as } n \to \infty.
\]
Hence using lemma 4.5, we conclude that

$$\sup_{\alpha \in I_n} |S_n^{lqr}(\alpha, \hat{G}_n) - S_n^{lqr}(\alpha, F_n)| \rightarrow 0, \quad \text{in } F_n\text{-probability.}$$

So, not only is the estimate $S_n^{lqr}(\alpha, \hat{G}_n)$ uniformly close to the asymptotic interquartile range $S^{lqr}(\alpha, F)$, but it is also uniformly close to the finite sample interquartile range $S_n^{lqr}(\alpha, F_n)$.

The next lemma is a general result about the convergence in probability of the argument minimizing a random function to the argument minimizing a fixed function whenever the random function converges uniformly in probability and the fixed function is continuous.

**Lemma 4.6:** Let $I$ and $I_n$ for $n = 1, 2, \ldots$ be subsets of $IR$. Suppose that $I_n$ is asymptotically dense for $I$, i.e., for each $x \in I$, given $\epsilon > 0$ there exists $N = N(x)$ such that for all $n > N$ there exists $x_n \in I_n$ such that $|x - x_n| < \epsilon$. Let $S(\alpha)$ be a fixed continuous function of $\alpha$ for all $\alpha \in I$, and suppose that

$$\sup_{\alpha \in I_n} |\hat{S}(\alpha) - S(\alpha)| \rightarrow 0, \quad \text{in probability},$$

where $\hat{S}(\cdot)$ is a random estimate $S(\cdot)$.

Let $\alpha_n = \arg\min_{\alpha \in I_n} \hat{S}(\alpha)$ and $A_0 = \arg\min_{\alpha \in I} S(\alpha)$, where we are assuming that $S(\alpha)$ has a unique global minimum over the closure of $I$. Then

$$\alpha_n \rightarrow A_0 \quad \text{in probability.}$$

**Proof:** This lemma is a slight generalization of lemma 3 in Jaeckel (1971). His proof is sufficient to prove this result.

Let $\alpha_n^{lqr}(\hat{G}_n) = \arg\min_{\alpha \in A_n} S_n^{lqr}(\alpha, \hat{G}_n)$ where $A_n = \{\alpha : \alpha \in I_n \text{ and } n\alpha \in N\}$ and $I_n = [\alpha_0, 1/2 - b_n]$. Applying this result to $\alpha_n^{lqr}$ gives the following lemma.
**Lemma 4.7:** Let \( F \) satisfy assumptions (A.1) through (A.3), and suppose that \( \{F_n\} \) satisfies condition (C). Assume that \( n^{1/2} b_n \to \infty \) as \( n \to \infty \). If \( b_n \to 0 \), let’s further assume that \( F \) satisfies assumption (A.4). Let \( \hat{G}_n \) be the empirical distribution function of a sample of size \( n \) from \( F_n \). Then

\[
\alpha_n^{lqr}(\hat{G}_n) \to A_0(F) \quad \text{in } F_n\text{-probability},
\]

where \( A_0(F) = \arg \inf_{\alpha \in I} S^{lqr}(\alpha, F) \) and \( I = [\alpha_0, 1/2) \) if \( b_n \to 0 \) or \( I = [\alpha_0, 1/2 - b_n] \) if \( b_n \) is a constant.

**Proof:** It is clear that \( A_n \) is asymptotically dense for \( I \). It is also clear that \( S^{lqr}(\alpha, F) \) is a continuous function in \( \alpha \) since \( \sigma^2(\alpha, F) \) is. Moreover, by assumption (A.3), it has only one global minimum. Finally, lemma 4.5 implies that the remaining condition of lemma 4.6 is also satisfied. Hence the conclusion of lemma 4.6 is the desired result.

**Remark 4.3:** As Jaeckel pointed out, assumption (A.3) can be relaxed. Basically, the idea consists of always taking the smallest value of \( \alpha \) such that \( S_n^{lqr}(\alpha, \hat{G}_n) \) is sufficiently close to \( \min_{\alpha \in I_n} S_n^{lqr}(\alpha, \hat{G}_n) \). To see this, first note that the result in lemma 4.3 could be strengthened to \( \sup_{\alpha \in I_n} |L_n(x; \alpha, F_n) - L(x; \alpha, F)| = O(n^{-1/2} b_n^{-1}) \) in \( F_n\)-probability. Thus with an argument such as the one in the proof of lemma 4.1, we could also strengthen lemma 4.5 to \( \sup_{\alpha \in I_n} |S_n^{lqr}(\alpha, \hat{G}_n) - S^{lqr}(\alpha, F)| = O(n^{-1/2} b_n^{-1}) \) in \( F_n\)-probability. So, let \( \{z_n\} \) be a sequence of positive integers such that \( z_n \to 0 \) while \( n^{1/2} b_n z_n \to \infty \). Let also \( A_0(F) = \inf\{\alpha \in [\alpha_0, 1/2) : S^{lqr}(\alpha, F) = \text{infinum}\} \). Then let \( M^2 = \min\{S_n^{lqr}(\alpha, \hat{G}_n) : \alpha \in I_n\} \). Finally, define \( \alpha_n^{lqr}(\hat{G}_n) \) to be the smallest \( \alpha \in I_n \) such that \( S_n^{lqr}(\alpha, \hat{G}_n) \leq M^2 + z_n \).

The reason that we can’t use \( z_n = 0 \) is that if there is more than one global minimum to \( S^{lqr}(\alpha, F) \) (say two), then the minimum of \( S_n^{lqr}(\alpha, \hat{G}_n) \) would be attained sometimes close to the smallest of the two minimums, at other times close to the largest. Hence there would therefore be no reason for the argument minimizing \( S_n^{lqr}(\alpha, \hat{G}_n) \) to converge. Instead, we let \( \alpha_n^{lqr}(\hat{G}_n) \) be the smallest value of \( \alpha \) such that \( S_n^{lqr}(\alpha, \hat{G}_n) \) is within \( z_n \) of the smallest of them. Since \( \sup_{\alpha \in I_n} |S_n^{lqr}(\alpha, \hat{G}_n) - S^{lqr}(\alpha, F)| = O(n^{-1/2} b_n^{-1}) \) in \( F_n\)-probability,
then for a given $\delta > 0$, there exists $N$ such that for $n > N$ with probability at least $1 - \delta$, $S_{n}^{\text{IQR}}(\alpha, \hat{G}_{n})$ is within a band of $O(n^{-1/2}b_{n}^{-1})$ of $S_{n}^{\text{IQR}}(\alpha, F)$ for all $\alpha \in I_{n}$. Therefore, $M^{2}$ must be within $O(n^{-1/2}b_{n}^{-1})$ of $S_{n}^{\text{IQR}}(A_{0}(F), F)$ which is itself within $O(n^{-1/2}b_{n}^{-1})$ of $S_{n}^{\text{IQR}}(A_{0}(F), \hat{G}_{n})$. Thus, $S_{n}^{\text{IQR}}(A_{0}(F), \hat{G}_{n})$ is within $O(n^{-1/2}b_{n}^{-1})$ of $M^{2}$. But since $z_{n}$ converges to 0 at a slower rate than $n^{-1/2}b_{n}^{-1}$, $S_{n}^{\text{IQR}}(A_{0}(F), \hat{G}_{n})$ is then smaller than $M^{2} + z_{n}$ for all large $n$, with probability at least $1 - \delta$. Thus, $\alpha_{n}^{\text{IQR}}(\hat{G}_{n}) \leq A_{0}(F)$ with probability at least $1 - \delta$, for all large $n$. Finally, an argument similar to that of lemma 3 in Jaeckel (1971) is sufficient to show that it cannot be much smaller than $A_{0}(F)$, and therefore that $\alpha_{n}^{\text{IQR}}(\hat{G}_{n}) \to A_{0}(F)$ in $F_{n}$-probability.

**Remark 4.4:** It is interesting to note that it is not necessary to assume that $A_{0}(F) \neq 1/2$ in this lemma, despite the fact that $I_{n}$ does not contain the value $1/2$ for any $n$.

We are almost ready to prove theorems 3.4 and 3.6. Let's first show that the distribution function of the bootstrap adaptive trimmed mean (using the interquartile range) based on a sample from $F_{n}$, $J_{n}^{\text{IQR}}(x, F_{n})$, converges to $\Phi(x/\sigma(A_{0}(F), F))$ uniformly in $x$ whenever $\{F_{n}\}$ satisfies condition (C).

**Lemma 4.8:** Let $F$ satisfy assumptions (A.1) through (A.3). If $b_{n} \to 0$, let's further assume that $F$ satisfies assumptions (A.4) and (A.5). Suppose that $\{F_{n}\}$ satisfies condition (C). Then

$$\sup_{x} |J_{n}^{\text{IQR}}(x, F_{n}) - \Phi(x/\sigma(A_{0}(F), F))| \to 0, \quad \text{as } n \to \infty.$$  

**Proof:** By lemma 4.7, $\alpha_{n}^{\text{IQR}}(\hat{G}_{n}) \to A_{0}(F)$ in $F_{n}$-probability. By assumption (A.5), $A_{0}(F) < 1/2$ and so by theorem 4.2, $M_{n}(\alpha, \hat{G}_{n}, F_{n})$ is $F_{n}$-stochastically equicontinuous at $A_{0}(F)$. Therefore, $M_{n}(\alpha_{n}^{\text{IQR}}, \hat{G}_{n}, F_{n})$ and $M_{n}(A_{0}(F), \hat{G}_{n}, F_{n})$ both have the same weak limit. But by corollary 4.1 $M_{n}(A_{0}(F), \hat{G}_{n}, F_{n})$ and $M_{n}(A_{0}(F), \hat{F}_{n}, F)$ also have the same weak limit, where $\hat{F}_{n}$ is the empirical distribution function of a sample of size $n$ from $F$. Also by corollary 2.2, the asymptotic distribution is Gaussian with asymptotic variance
equal to $\sigma^2(A_0(F), F)$. Therefore, for all $x$,

$$|J_{n}^{\text{lr}}(x, F_n) - \Phi(x/\sigma(A_0(F), F))| \to 0, \quad \text{as } n \to \infty.$$ 

But since $\Phi(\cdot)$ is a continuous function, we conclude that

$$\sup_x |J_{n}^{\text{lr}}(x, F_n) - \Phi(x/\sigma(A_0(F), F))| \to 0, \quad \text{as } n \to \infty.$$ 

\[ \Box \]

**Proof of Theorem 3.4:** The proof is now immediate: let $F_n = F$ for all $n$ in lemma 4.8. \[ \Box \]

**Proof of Theorem 3.6:** By lemma 4.4, there exists $\{\hat{F}_n\}$ with the same distribution as $\{\hat{F}_n\}$ for which condition (C) is satisfied. Hence lemma 4.8 implies

$$\sup_x |J_{n}^{\text{lr}}(x, \hat{F}_n) - \Phi(x/\sigma(A_0(F), F))| \to 0, \quad \text{almost surely.}$$

But since $\{\hat{F}_n\}$ and $\{\hat{F}_n\}$ have the same distribution, we conclude that

$$\sup_x |J_{n}^{\text{lr}}(x, \hat{F}_n) - \Phi(x/\sigma(A_0(F), F))| \to 0, \quad \text{in probability.} \quad (4.29)$$

Also by theorem 3.4,

$$\sup_x |J_{n}^{\text{lr}}(x, F) - \Phi(x/\sigma(A_0(F), F))| \to 0.$$

Hence,

$$\sup_x |J_{n}^{\text{lr}}(x, \hat{F}_n) - J_{n}^{\text{lr}}(x, F)| \to 0, \quad \text{in probability.}$$

\[ \Box \]

Let us now show that theorem 3.6 is sufficient to imply that a bootstrap confidence interval based on the bootstrap adaptive trimmed mean has asymptotically the claimed coverage probability.
Proof of Corollary 3.3: Consider the one-sided bootstrap confidence interval given by
\[ \theta \in [-\infty, T_n^{lqr}(\hat{F}_n) - n^{-1/2} z_{\beta}(\hat{F}_n)], \]
where \( z_{\beta}(\hat{F}_n) = J_n^{lqr-1}(\beta, \hat{F}_n) \). Then,
\[
Prob_F \left\{ \theta \in [-\infty, T_n^{lqr}(\hat{F}_n) - n^{-1/2} z_{\beta}(\hat{F}_n)] \right\} = Prob_F \left\{ \theta \leq T_n^{lqr}(\hat{F}_n) - n^{-1/2} z_{\beta}(\hat{F}_n) \right\}
= Prob_F \left\{ M_n(\alpha_n^{lqr}(\hat{F}_n), \hat{F}_n, F) \geq J_n^{lqr-1}(\beta, \hat{F}_n) \right\}
= 1 - J_n^{lqr}(J_n^{lqr-1}(\beta, \hat{F}_n) - 0, F),
\]
by definition of \( z_{\beta}(\hat{F}_n) \), where \( g(x - 0) = \lim_{y \to x^-} g(y) \). But by (4.29), the continuity of \( \Phi \) and an argument like that in lemma 4.1,
\[
J_n^{lqr-1}(\beta, \hat{F}_n) \to \sigma(A_0(F), F) \Phi^{-1}(\beta), \quad \text{in probability.}
\]
This result and (4.29) implies that
\[
J_n^{lqr}(J_n^{lqr-1}(\beta, \hat{F}_n) - 0, F) \to \beta.
\]
Hence,
\[
\lim_{n \to \infty} Prob_F \left\{ \theta \in [-\infty, T_n^{lqr}(\hat{F}_n) - n^{-1/2} z_{\beta}(\hat{F}_n)] \right\} = 1 - \beta.
\]
Consequently, the asymptotic coverage probability of the two-sided bootstrap confidence interval given by (3.13) is \( 1 - 2\beta \), as claimed.

Remark 4.5: Note that this bootstrap confidence interval involves a double bootstrap. This method of proof is so flexible that it could handle any level of bootstrapping. For instance, suppose that we want to show that
\[
\sup_x |J_n^{lqr}(x, \hat{G}_n) - \Phi(x/\sigma(A_0(F)))| \to 0, \quad \text{in probability,}
\]
where \( \hat{G}_n \) is the empirical distribution function of a sample of size \( n \) from \( \hat{F}_n \) (conditional on its observed value), the empirical distribution function from a sample of size \( n \) from \( F \). Applying lemma 4.4 twice, there exists \( \{\hat{G}_n\} \) with the same distribution as \( \{\hat{G}_n\} \) as
defined above for which condition (C) is satisfied. Hence, one can apply lemma 4.8 with $F_n = \hat{G}_n$ and then replace $\hat{G}_n$ by $\hat{G}_n$ to get the desired result.

This completes our discussion of the bootstrap adaptive trimmed mean based on the interquartile range. The bulk of the work has now been done. To prove the results about the bootstrap adaptive trimmed mean based on the finite sample variance and Jaeckel’s estimate, we will follow the same outline as above. The only results that will require new proofs will be the lemmas corresponding to lemma 4.3. The other corresponding results will not need any new proofs.

Let’s now focus on the bootstrap adaptive trimmed mean based on the finite sample variance. The development will parallel that of the interquartile range.

It will first be shown that $S_n^{\text{Var}}(\alpha, F_n)$ is uniformly close to $S^{\text{Var}}(\alpha, F) = \sigma^2(\alpha, F)$ for $\sigma^2(\alpha, F)$ as defined in (2.2), whenever $\{F_n\}$ satisfies condition (C).

**Lemma 4.9:** Suppose that $F$ satisfies (A.1) and (A.2) and that $\{F_n\}$ satisfies condition (C). Let $n^{1/2} b_n \to \infty$ as $n \to \infty$. Then

$$\sup_{\alpha \in \mathcal{I}_n} |S_n^{\text{Var}}(\alpha, F_n) - S^{\text{Var}}(\alpha, F)| \to 0,$$

where $\mathcal{I}_n$ is the interval $[\alpha_0, 1/2 - b_n]$.

**Proof:** From decomposition (4.10) of lemma 4.2,

$$S_n^{\text{Var}}(\alpha, F_n) = v(\alpha, F_n) + \text{Var}(n^{1/2} \Delta_n(\alpha, \hat{G}_n, F_n))$$

$$+ 2 \text{Cov}(n^{1/2} dM_n(\alpha, \hat{G}_n, F_n), n^{1/2} \Delta_n(\alpha, \hat{G}_n, F_n)).$$

If we can show that the variance term converges to 0 uniformly in $\alpha$, then so will the covariance term, and so we would conclude that

$$\sup_{\alpha \in \mathcal{I}_n} |S_n^{\text{Var}}(\alpha, F_n) - v(\alpha, F_n)| \to 0, \quad \text{as } n \to \infty.$$

Therefore

$$\lim_{n \to \infty} \sup_{\alpha \in \mathcal{I}_n} |S_n^{\text{Var}}(\alpha, F_n) - S^{\text{Var}}(\alpha, F)| = \lim_{n \to \infty} \sup_{\alpha \in \mathcal{I}_n} |v(\alpha, F_n) - \sigma^2(\alpha, F)|$$

$$= 0,$$
Section 4.2: Proofs

by (4.22).

So let's show that the variance term converges to 0 uniformly in $\alpha$. It is sufficient to show that

$$E \left[ n^{1/2} \sup_{\alpha \in \Omega} \Delta_n(\alpha, \hat{G}_n, F_n) \right]^2 \to 0. \quad (4.30)$$

By (4.25), we have

$$E \left[ n^{1/2} \sup_{\alpha \in \Omega} \Delta_n(\alpha, \hat{G}_n, F_n)^4 \right] \leq E \left( \sup_{\alpha \in \Omega} ||K_\alpha(\cdot; \hat{G}_n, F_n)||_{L_1} \right)^4 \left[ n^{1/2} ||\hat{G}_n - F_n||_\infty \right]^4
\leq \sqrt{E \left[ \sup_{\alpha \in \Omega} ||K_\alpha(\cdot; \hat{G}_n, F_n)||_{L_1} \right]^8 E\left[ n^{1/2} ||\hat{G}_n - F_n||_\infty \right]^8}, \quad (4.31)$$

by Schwarz's inequality.

By DKW's inequality,

$$\text{Prob}_{F_n}\{n^{1/2}||\hat{G}_n - F_n||_\infty > t\} \leq C \exp\{-t^2\},$$

hence,

$$\text{Prob}_{F_n}\{[n^{1/2}||\hat{G}_n - F_n||_\infty]^8 > t\} \leq C \exp\{-t^{4/8}\}. \quad (4.32)$$

So, using the well-known equality that if $X$ is a positive random variable then $E[X] = \int_0^\infty \text{Prob}[X > t] \, dt$, we have

$$E[n^{1/2}||\hat{G}_n - F_n||_\infty]^8 = \int_0^\infty \text{Prob}_{F_n}\{[n^{1/2}||\hat{G}_n - F_n||_\infty]^8 > t\} \, dt \leq \int_0^\infty C \exp\{-t^{4/8}\} \, dt < \infty, \quad (4.32)$$

uniformly in $n$.

Likewise, from (4.26) we have

$$\text{Prob}_{F_n}\left\{ \sup_{\alpha \in \Omega} ||K_\alpha(\cdot; \hat{G}_n, F_n)||_{L_1} \geq 8b_n^{-1}n^{-1/2}M_2/f_0 \right\} \leq C \exp\{-M_1^2\},$$

where $M_2 = 2 \max(D, M_1)$. Hence

$$\text{Prob}_{F_n}\left\{ \sup_{\alpha \in \Omega} ||K_\alpha(\cdot; \hat{G}_n, F_n)||_{L_1} \geq t \right\} \leq C \exp\{-[b_n^{n^{1/2}}f_0^2/16]^2\}, \quad (4.33)$$
for all $t > 16b_n^{-1}n^{-1/2}D/f_0$. But since $n^{1/2}b_n \to \infty$, the sequence $n^{-1/2}b_n^{-1}$ is bounded above and so (4.33) is valid for all $n$ and all $t > C_1$ for some constant $C_1$. Therefore

$$E \left[ \sup_{\alpha \in \mathcal{F}_n} \| K_\alpha (\cdot ; \hat{G}_n, F_n) \|_{L_1} \right]^8 = \int_0^\infty \text{Prob} \left\{ \left[ \sup_{\alpha \in \mathcal{F}_n} \| K_\alpha (\cdot ; \hat{G}_n, F_n) \|_{L_1} \right]^8 \geq t \right\} dt \leq C_1 + C \int_{C_1}^\infty \exp \left\{ - \left[ b_n n^{1/2} f_0 t / 16 \right]^{1/4} \right\} dt < \infty, \quad (4.34)$$

uniformly in $n$. Using (4.34) and (4.32), we see that (4.31) is bounded above uniformly in $n$. Hence

$$\sup_{\alpha \in \mathcal{F}_n} E \left[ n^{1/2} \sup_{\alpha \in \mathcal{F}_n} \Delta_n (\alpha, \hat{G}_n, F_n) \right]^4 < \infty.$$ 

This implies (4.30) since $\sup_{\alpha \in \mathcal{F}_n} \Delta_n (\alpha, \hat{G}_n, F_n) \to 0$ in $F_n$-probability and therefore weakly.

\textbf{Remark 4.6:} Note that, unlike in lemma 4.2, the density is not assumed to be bounded at the median. Recall that this was needed to guarantee (4.24), i.e., that the difference between the $v(\alpha, F_n)$ and $\sigma^2(\alpha, F)$ divided by $v(\alpha, F_n)\sigma^2(\alpha, F)$ did converge to 0, and so $\sigma^2(\alpha, F)$ better not be 0.

This result is interesting in that it does not require any other conditions on $F$ to guarantee that not only the distribution functions of the trimmed mean converge uniformly in $\alpha$, but also that the respective second moments also converge uniformly in $\alpha$. Moreover, conditions (A.1) and (A.2) do not put any restrictions on the moments of $F$. Note that the corresponding result for the median does not hold. Ghosh, Parr, Singh, and Babu (1984) showed that if $E|X^a| < \infty$ for some $a > 0$ where $X$ has distribution $F$, then the bootstrap variance converges to $\sigma^2(\alpha, F)$ almost surely. On the other hand, they have provided a counterexample to that result when the condition is not satisfied.

\textbf{Lemma 4.10:} Suppose that $F$ satisfies (A.1) and (A.2) and that $\{F_n\}$ satisfies condition (C). Let $n^{1/2}b_n \to \infty$ as $n \to \infty$. Then

$$\sup_{\alpha \in \mathcal{F}_n} |S_n^{\text{Var}}(\alpha, \hat{G}_n) - S^{\text{Var}}(\alpha, F)| \to 0 \quad \text{in } F_n\text{-probability},$$
where \( I_n \) is the interval \([a_0, 1/2 - b_n]\).

**Proof:** Same as in lemma 4.5. \( \square \)

**Lemma 4.11:** Suppose that \( F \) satisfies (A.1) through (A.3) and that \( \{F_n\} \) satisfies condition (C). Let \( n^{1/2}b_n \to \infty \) as \( n \to \infty \). Then

\[
\alpha_n^{\text{Var}}(\hat{G}_n) \to A_0(F), \quad \text{in } F_n\text{-probability.}
\]

**Proof:** As in lemma 4.7. \( \square \)

**Lemma 4.12:** Suppose that \( F \) satisfies (A.1) through (A.3) and that \( \{F_n\} \) satisfies condition (C). Let \( n^{1/2}b_n \to \infty \) as \( n \to \infty \). Then

\[
\sup_x |J_n^{\text{Var}}(x, F_n) - \Phi(x/\sigma(A_0(F), F))| \to 0, \quad \text{as } n \to \infty.
\]

**Proof:** Same as lemma 4.8. \( \square \)

To prove theorems 3.3 and 3.5, follow the steps of the proofs of theorems 3.4 and 3.6.

Let's now look at Jaeckel's estimate. First, it is shown that \( S^{\text{Jae}}(\alpha, F_n) \) is uniformly close to \( S^{\text{Jae}}(\alpha, F) \) whenever \( \{F_n\} \) satisfies condition (C).

**Lemma 4.13:** Suppose that \( F \) satisfies (A.1) and (A.2) and that \( \{F_n\} \) satisfies condition (C). Let \( n^{1/4}b_n \to \infty \) as \( n \to \infty \). Then

\[
\sup_{\alpha \in I_n} |S^{\text{Jae}}(\alpha, F_n) - S^{\text{Jae}}(\alpha, F)| \to 0,
\]

where \( I_n \) is the interval \([a_0, 1/2 - b_n]\).
Proof: Remember that $S_{\text{je}}(\alpha, F_n) = \sigma^2(\alpha, F_n)$ where $\sigma^2(\alpha, F)$ is defined in (2.2). Through integration by parts,

$$\sigma^2(\alpha, F) = (1 - 2\alpha)^{-2} \left[ \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [x - F^{-1}(1/2)]^2 dF(x) + \alpha(F^{-1}(\alpha) - F^{-1}(1/2))^2 ight. \\
\left. + \alpha(F^{-1}(1 - \alpha) - F^{-1}(1/2))^2 \right]$$

$$= (1 - 2\alpha)^{-2} \left[ -2 \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [x - F^{-1}(1/2)]F(x) dx \\
+ [F^{-1}(1 - \alpha) - F^{-1}(1/2)]^2[F(F^{-1}(1 - \alpha)) + \alpha] \\
+ [F^{-1}(\alpha) - F^{-1}(1/2)]^2[\alpha - F(F^{-1}(1/2))] \right]$$

$$= (1 - 2\alpha)^{-2}[T_1(\alpha, F) + T_2(\alpha, F) + T_3(\alpha, F)].$$

Hence,

$$|\sigma^2(\alpha, F) - \sigma^2(\alpha, F_n)| \leq (1 - 2\alpha)^{-2} \sum_{i=1}^{3} |T_i(\alpha, F) - T_i(\alpha, F_n)|. \quad (4.35)$$

Clearly, conditions (C) and (C*) imply that

$$\sup_{\alpha \in I_n} |T_i(\alpha, F) - T_i(\alpha, F_n)| = O(n^{-1/2}), \quad \text{for } i = 2, 3.$$ 

Also,

$$|T_1(\alpha, F) - T_1(\alpha, F_n)| = 2 \left| \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [x - F^{-1}(1/2)]F(x) dx \\
- \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [x - F_n^{-1}(1/2)]F_n(x) dx \right|$$

$$\leq 2 \left| \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [x - F^{-1}(1/2)]F(x) dx \\
- \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [x - F_n^{-1}(1/2)]F_n(x) dx \right|$$

$$+ 2 \left| \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} [x - F_n^{-1}(1/2)]F_n(x) dx \\
- \int_{F^{-1}(\alpha)}^{F_n^{-1}(1-\alpha)} [x - F_n^{-1}(1/2)]F_n(x) dx \right|$$

$$= 2(U_n^1(\alpha) + U_n^2(\alpha)). \quad (4.36)$$
Now,
\[
\begin{align*}
\sup_{\alpha \in I_n} U_n^1(\alpha) & \leq \sup_{\alpha \in I_n} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \left| x[F(x) - F_n(x)] + [F_n^{-1}(1/2)F_n(x) - F^{-1}(1/2)F(x)] \right| dx \\
& = O(n^{-1/2}),
\end{align*}
\]
(4.37)

by conditions (C) and (C*). Finally,
\[
\begin{align*}
\sup_{\alpha \in I_n} U_n^2(\alpha) & \leq C \text{ Leb}(S_n(\alpha)) \\
& = O(n^{-1/2}),
\end{align*}
\]
(4.38)

for a fixed constant C and where the set \(S_n(\alpha)\) is defined in (4.18). Hence using (4.37) and (4.38) in (4.36) gives
\[
\sup_{\alpha \in I_n} |T_1(\alpha, F) - T_1(\alpha, F_n)| = O(n^{-1/2}).
\]

Going back to (4.35), we get
\[
\sup_{\alpha \in I_n} |\sigma^2(\alpha, F) - \sigma^2(\alpha, F_n)| = O(n^{-1/2}b_n^{-2}).
\]

And so the result obtains provided that \(n^{1/4}b_n \to \infty\).
\[\blacksquare\]

**Remark 4.7:** As was mentioned in chapter 3, the assumption that \(n^{1/4}b_n \to \infty\) is different from \(n^{1/2}b_n\) assumed in lemma 4.2. It seems that this assumption is more than an artifact of the method of proof. This same assumption was also required when Jaekel's original proof was adapted. Note also that assumption (A.4) is no longer required when \(b_n \to 0\).

Proceeding as in lemmas 4.5, 4.7, and 4.8, we have the following corresponding results for Jaekel's estimator.

**Lemma 4.14:** Suppose that \(F\) satisfies (A.1) and (A.2) and that \(\{F_n\}\) satisfies condition (C). Let \(n^{1/4}b_n \to \infty\) as \(n \to \infty\). Then
\[
\sup_{\alpha \in I_n} |S^{\text{jae}}(\alpha, \hat{G}_n) - S^{\text{jae}}(\alpha, F)| \to 0 \quad \text{in } F_n\text{-probability}.
\]
where $I_n$ is the interval $[\alpha_0, 1/2 - b_n]$.

**Proof:** Same as in lemma 4.5.

**Lemma 4.15:** Suppose that $F$ satisfies (A.1) through (A.3) and that $\{F_n\}$ satisfies condition (C). Let $n^{1/4}b_n \to \infty$ as $n \to \infty$. Then

$$\alpha_{n}^{\text{Jae}}(\hat{G}_n) \to A_0(F), \quad \text{in } F_n\text{-probability}.$$

**Proof:** As in lemma 4.7.

**Lemma 4.16:** Suppose that $F$ satisfies (A.1) through (A.3) and that $\{F_n\}$ satisfies condition (C). Let $n^{1/4}b_n \to \infty$ as $n \to \infty$. Then

$$\sup_x |J_n^{\text{Jae}}(x, F_n) - \Phi\{x/\sigma(A_0(F), F)\}| \to 0, \quad \text{as } n \to \infty.$$

**Proof:** Same as lemma 4.8.

To prove theorems 3.1 and 3.2, follow the steps of the proofs of theorems 3.4 and 3.6. This completes out treatment of Jaeckel's estimator.

All of the results that have so far been proven apply to the usual bootstrap, i.e. when the resampling is done from the empirical distribution function. As should now become clear, the flexibility of the method of proof will allow us to prove that all of those results also apply to the smooth and symmetric bootstrap as stated in theorems 3.7 and 3.8. Clearly, only lemma 4.4 needs to be modified to prove those results. Let's start with the smooth bootstrap.

**Lemma 4.17:** Let $\{F_n\}$ satisfy condition (C) for some distribution function $F$ with bounded density $f$. Let $Y_{n,1}, \ldots, Y_{n,n}$ be a sample of size $n$ from $F_n$. Let $\hat{G}_{n,c_n}$ be the kernel distribution function estimate based on the $Y$'s and distribution function $K$, as defined in (3.14). Then if $n^{1/2}c_n \to 0$ as $n \to \infty$ and $K$ has a finite first moment,
there exists \( \{ \tilde{G}_n \} \) with the same distribution as \( \{ \hat{G}_{n,c_n} \} \) such that it satisfies condition (C) almost surely.

The same result holds if \( n^{1/2} c_n^2 \to 0 \), under the stronger assumptions that \( f \) is differentiable with a uniformly bounded derivative, and that \( K \) has a finite second moment.

**Proof:** As in lemma 4.4, let \( \{ \xi_i \} \) be a sequence of independent random variables each having a uniform distribution on the interval \((0, 1)\) such that \( \| U_n - U \|_\infty \to 0 \) almost surely where \( \|x\|_\infty = \sup_r |x(r)| \), \( U_n(t) = n^{-1/2} \sum_{i=1}^n [I\{ \xi_i \leq t \} - t] \) for all \( 0 \leq t \leq 1 \) and \( U \) is a Brownian bridge.

Let \( Y_{n,1}, \ldots, Y_{n,n} \) be a sample of size \( n \) from \( F_n \). Then consider the process \( X_n^*(x) = n^{1/2}[\hat{G}_{n,c_n}(x) - F_n(x)] \). Through integration by parts, \( X_n^* \) can be rewritten as follows,

\[
X_n^*(x) = n^{1/2}[\hat{G}_{n,c_n}(x) - E\hat{G}_{n,c_n}(x) + E\hat{G}_{n,c_n}(x) - F_n(x)]
\]

\[
= \int_{-\infty}^{\infty} K \left( \frac{x - y}{c_n} \right) dU_n^*(F_n(y)) + n^{1/2}(E\hat{G}_{n,c_n}(x) - F_n(x))
\]

\[
= -\int_{-\infty}^{\infty} U_n^*(F_n(y)) dK \left( \frac{x - y}{c_n} \right) + n^{1/2}(E\hat{G}_{n,c_n}(x) - F_n(x))
\]

\[
= -\int_{-\infty}^{\infty} U_n^*(F_n(x - sc_n)) dK(s) + n^{1/2}(E\hat{G}_{n,c_n}(x) - F_n(x)), \quad (4.39)
\]

where \( U_n^*(x) = n^{-1/2} \sum_{i=1}^n [I\{ F_n(Y_{n,i}) \leq x \} - x] \). But \( U_n \) as defined above has the same distribution as \( U_n^* \). Hence letting

\[
\tilde{G}_n(x) = \frac{1}{n} \sum_{i=1}^n K \left( \frac{x - F_n^{-1}(\xi_i)}{c_n} \right),
\]

and defining \( X_n = n^{1/2}[\tilde{G}_n(x) - F_n(x)] \), we see at once that (4.39) implies that

\[
X_n(x) = -\int_{-\infty}^{\infty} U_n(F_n(x - sc_n)) dK(s) + n^{1/2}(E\hat{G}_{n,c_n}(x) - F_n(x)).
\]

Therefore,

\[
X_n(x) - U(F(x)) = -\int_{-\infty}^{\infty} [U_n(F_n(x - sc_n)) - U(F(x))] dK(s) + n^{1/2}(E\hat{G}_{n,c_n}(x) - F_n(x))
\]

\[
= R_n^1(x) + R_n^2(x), \quad (4.40)
\]

for \( U_n \) and \( U \) as defined above.
Hence

\[ \| R^1_n \|_\infty = \sup_x |R^1_n(x)| \]
\[ \leq \sup_x \int_{-\infty}^{\infty} |U_n(F_n(x - sc_n)) - U(F(x))| \, dK(s) \]
\[ \leq \sup_x \int_{-\infty}^{\infty} |U_n(F_n(x - sc_n)) - U(F_n(x - sc_n))| \, dK(s) \]
\[ + \sup_x \int_{-\infty}^{\infty} |U(F_n(x - sc_n)) - U(F(x))| \, dK(s) \]
\[ + \sup_x \int_{-\infty}^{\infty} |U(F(x - sc_n)) - U(F(x))| \, dK(s) \]
\[ \leq \| U_n - U \|_\infty + \| T^1_n \|_\infty + \| T^2_n \|_\infty. \] (4.41)

The first term in (4.41) converges to 0 almost surely, by the special construction of those processes. Since \( U \) is uniformly continuous and \( \{ F_n \} \) satisfies condition (C), \( \| T^1_n \|_\infty \) also converges to 0 almost surely. The last term requires somewhat more care.

Let \( \epsilon > 0 \) be given. Since \( U \) is uniformly continuous, there exists \( \delta > 0 \) such that

\[ |U(x) - U(y)| < \epsilon/2 \]

provided that \( |x - y| < \delta \). But since \( F \) is also uniformly continuous, there exists \( \gamma > 0 \) such that \( |F(x) - F(y)| < \delta \) provided that \( |x - y| < \gamma \). Moreover, since \( U \) is continuous on a compact set, it is bounded, say by \( M \). Therefore, let \( A \) be chosen in such a way that \( \text{Prob}_K([-A, A]^C) < \epsilon/(4M) \). Then, since \( c_n \to 0 \), there exists \( N \) such that \( Ac_n < \gamma \) for all \( n > N \). Thus

\[ T^2_n(x) = \int_{-\infty}^{\infty} |U(F(x - sc_n)) - U(F(x))| \, dK(s) \]
\[ = \int_{s \in [-A, A]} |U(F(x - sc_n)) - U(F(x))| \, dK(s) \]
\[ + \int_{s \notin [-A, A]} |U(F(x - sc_n)) - U(F(x))| \, dK(s) \]
\[ \leq \int_{-A}^{A} (\epsilon/2) \, dK(s) + 2M \text{Prob}_K([-A, A]^C) \]
\[ < (\epsilon/2) + 2M \epsilon/(4M) = \epsilon. \]

Hence \( \| R^1_n \|_\infty \) converges to 0 almost surely.
Finally,

$$|R_n^2(x)| = n^{1/2}\left|E\hat{C}_{n,c_n}(x) - F_n(x)\right|$$

$$= n^{1/2}\left|\int_{-\infty}^{\infty} K\left(\frac{x-y}{c_n}\right) dF_n(y) - F_n(x)\right|$$

$$= n^{1/2}\left|\int_{-\infty}^{\infty} [F_n(x - sc_n) - F_n(x)] dK(s)\right|$$

$$\leq n^{1/2}\left|\int_{-\infty}^{\infty} [F_n(x - sc_n) - F(x - sc_n)] + [F(x) - F_n(x)] dK(s)\right|$$

$$+ n^{1/2}\left|\int_{-\infty}^{\infty} [F(x - sc_n) - F(x)] dK(s)\right|$$

$$\leq 2D + n^{1/2}c_n \left|\int_{-\infty}^{\infty} s f(x_s) dK(s)\right|,$$

where $x_s$ is between $x$ and $x - sc_n$, and using condition (C). But by assumption $f$ is bounded. Therefore,

$$\|R_n^2\|_\infty \leq 2D + n^{1/2}c_n \|f\|_\infty \int_{-\infty}^{\infty} |s| dK(s).$$

Again, by assumption $n^{1/2}c_n \to 0$ as $n \to \infty$ and so the right hand side converges to $2D$ since $K$ has a first moment.

If instead, it is assumed that $f$ is differentiable, that $f'$ is uniformly bounded and that $K$ has a second moment, then we can use the following approximation,

$$n^{1/2}\left|\int_{-\infty}^{\infty} [F(x - sc_n) - F(x)] dK(s)\right| = n^{1/2}c_n^2 \int_{-\infty}^{\infty} s^2 f(x_s) dK(s)$$

$$\leq n^{1/2}c_n^2 \|f'\|_\infty \int_{-\infty}^{\infty} s^2 dK(s).$$

In this case, the weaker assumption that $n^{1/2}c_n^2 \to 0$ is sufficient.

In both instances,

$$\sup_x |n^{1/2}[\tilde{C}_n(x) - F_n(x)] - U(F(x))| \leq 2D, \quad \text{almost surely.}$$

Thus, as in lemma 4.4, for almost every sample path of $\tilde{G}_n$, there exists $2D \leq D_1 < \infty$ such that for all large $n$,

$$n^{1/2} \sup_x |\tilde{G}_n(x) - F_n(x)| \leq D_1.$$
Moreover, by condition (C)

\[ n^{1/2} \sup_r \left| F_n(r) - F(r) \right| \leq D, \]

for all large \( n \). Consequently, for almost every sample paths of \( \tilde{G}_n \), there exists \( D_2 = D_1 + D \) such that for all large \( n \),

\[ n^{1/2} \sup_r \left| \tilde{G}_n(r) - F(r) \right| \leq D, \]

i.e., \( \{\tilde{G}_n\} \) satisfies condition (C) almost surely.

The proof of theorem 3.7 follows at once by replacing all references to lemma 4.4 in lemmas 4.5, 4.7, and 4.8 by lemma 4.17.

\textbf{Lemma 4.18:} Let \( \{F_n\} \) satisfy condition (C) with \( F \) being symmetric about \( \theta \), and having a bounded density \( f \). Assume also that \( F_n \) is symmetric about \( \theta_n \) where \( n^{1/2}[\theta_n - \theta] = O(1) \). Let \( Y_{n,1}, \ldots, Y_{n,n} \) be a sample of size \( n \) from \( F_n \), and let \( \hat{G}_n(x) = (1/2)[\hat{G}_n(x) + 1 - \hat{G}_n(2\hat{G}_n^{-1}(1/2) - x - 0)] \), where \( \hat{G}_n \) is the empirical distribution function, \( \hat{G}_n^{-1}(1/2) \) is the median, \( \hat{G}_n(x - 0) = \lim_{y \to x^-} \hat{G}_n(y) \). Then there exists \( \{\hat{G}_n\} \) with the same distribution as \( \{\tilde{G}_n\} \) such that it satisfies condition (C) almost surely. Also \( n^{1/2}[\hat{G}_n^{-1}(1/2) - \theta] = O(1) \) almost surely.

\textbf{Proof:} As in lemma 4.4, let \( \{\xi_i\} \) be a sequence of independent random variables each having a uniform distribution on the interval \( (0, 1) \) such that \( \|U_n - U\| \to 0 \) almost surely where \( \|x\| = \sup_r |x(r)| \), \( U_n(t) = n^{1/2}[\tilde{H}_n(t) - t] \) for all \( 0 \leq t \leq 1 \), \( \tilde{H}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I\{\xi_i \leq t\} \), and \( U \) is a Brownian bridge.

Let \( \tilde{G}_n(x) = \tilde{H}_n(F_n(x)) \). Then \( \tilde{G}_n \) has the same distribution as \( \hat{G}_n \), the empirical distribution of a sample of size \( n \) from \( F_n \). Moreover, as in lemma 4.4,

\[ \sup_x \left| n^{1/2} [\tilde{G}_n(x) - F_n(x)] - U(F_n(x)) \right| \to 0, \quad \text{almost surely}. \quad (4.42) \]

Furthermore, Theorem 3.1.1 of Shorack and Wellner also implies that

\[ \sup_x |V_n(x) - V(x)| \to 0, \quad \text{almost surely}, \]
where \( V_n(x) = n^{1/2}[\hat{H}_n^{-1}(x) - x] \) and \( V = -U \) is a Brownian bridge.

Note that \( \tilde{G}_n^{-1}(x) = F_n^{-1}(\hat{H}_n^{-1}(x)) \). Hence

\[
|n^{1/2}[\tilde{G}_n^{-1}(1/2) - F_n^{-1}(1/2)] - F_n^{-1}(V(1/2))| = |F_n^{-1}(V_n(1/2)) - F_n^{-1}(V(1/2))| \\
\rightarrow 0, \quad \text{almost surely.} \quad (4.43)
\]

To prove the convergence in (4.43), note that by condition (C), \( F_n \) converges weakly to \( F \) which implies that \( F_n^{-1}(x) \rightarrow F^{-1}(x) \) for all continuity points \( x \) of \( F^{-1} \). But \( F^{-1} \) has at most countably many discontinuities. So the set of discontinuities has Lebesgue measure zero, and since \( V(1/2) \) has a normal distribution with mean 0 and variance 1/4, the probability that \( V(1/2) \) is in that set is 0. Hence the convergence of \( F_n^{-1} \) to \( F^{-1} \), and the almost sure convergence of \( V_n(1/2) \) to \( V(1/2) \) both imply the convergence in (4.43).

Let \( \hat{\theta}_n = \tilde{G}_n^{-1}(1/2) \). Then let \( \check{G}_n(x) = 1/2[\tilde{G}_n(x) + 1 - \tilde{G}_n(2\hat{\theta}_n - x - 0)] \) where \( k(x - 0) = \lim_{y \to x^-} k(y) \). That symmetric estimate has the same distribution as \( \tilde{G}_n \).
Without loss of generality, assume that \( F \) is symmetric about 0. Then

\[
n^{1/2}[\check{G}_n(x) - F(x)] = n^{1/2}\left\{ (1/2)[\tilde{G}_n(x) + 1 - \tilde{G}_n(\hat{\theta}_n - x - 0)] - F(x) \right\} \\
= (1/2)n^{1/2}\left\{ \tilde{G}_n(x) - F_n(x) - \tilde{G}_n(2\hat{\theta}_n - x - 0) + F_n(2\hat{\theta}_n - x - 0) \right\} \\
+ F_n(x) - F_n(2\hat{\theta}_n - x - 0) + 1 - 2F(x) \\
= (1/2)n^{1/2}\left\{ [\tilde{G}_n(x) - F_n(x)] - [\tilde{G}_n(2\hat{\theta}_n - x - 0) - F_n(2\hat{\theta}_n - x - 0)] \right\} \\
+ [F_n(x) - F(x)] - [F_n(2\hat{\theta}_n - x - 0) - F(-x)],
\]

since \( F(x) + F(-x) = 1 \) by the assumed symmetry of \( F \) about 0. Hence,

\[
n^{1/2}[\check{G}_n(x) - F(x)] = (1/2)\left\{ n^{1/2}[\tilde{G}_n(x) - F_n(x)] \right\} \\
- (1/2)\left\{ n^{1/2}[\tilde{G}_n(2\hat{\theta}_n - x - 0) - F_n(2\hat{\theta}_n - x - 0)] \right\} \\
+ (1/2)n^{1/2}\left\{ F_n(x) - F(x) \right\} \\
- (1/2)n^{1/2}\left\{ F_n(2\hat{\theta}_n - x - 0) - F(-x) \right\} \\
= T_1^n(x) + T_2^n(x) + T_3^n(x) + T_4^n(x).
\]

Note that by (4.42) and the almost sure continuity of \( U \) on the compact space \([0,1],\)
both $\|T^1_n\|_\infty$ and $\|T^2_n\|_\infty$ are $O(1)$ almost surely. Moreover, by condition (C), $\|T^3_n\|_\infty \leq (1/2)D$. Finally,

$$|T^4_n(x)| \leq (1/2)n^{1/2}|F_n(2\hat{\theta}_n - x - 0) - F(2\hat{\theta}_n - x)| + (1/2)n^{1/2}|F(2\hat{\theta}_n - x) - F(-x)|$$
$$\leq (1/2)D + n^{1/2}[\hat{\theta}_n - \theta_n]f(x^*) + n^{1/2}\theta_n f(x^*),$$

where $x^*$ is between $-x$ and $2\hat{\theta}_n - x$, and where we have used the assumption that $\theta = 0$. But by assumption both $f$ and $n^{1/2}\theta_n$ are bounded. Moreover, (4.43) implies that $n^{1/2}[\hat{\theta}_n - \theta_n] = n^{1/2}[\tilde{G}^{-1}_n(1/2) - \theta_n]$ is bounded almost surely. Thus, $\|T^4_n\|_\infty$ is bounded almost surely, which in turn implies that

$$n^{1/2}\sup_{x}|\tilde{G}_n(x) - F(x)|$$

is bounded almost surely, i.e., $\{\tilde{G}_n\}$ satisfies condition (C) with probability one.

Note also that $\tilde{G}_n$ is symmetric about $\tilde{G}^{-1}_n(1/2)$. Hence $\tilde{G}^{-1}_n(1/2) = \tilde{G}^{-1}_n(1/2)$, and so as noted above we conclude that $n^{1/2}[\tilde{G}^{-1}_n(1/2) - \theta] = O(1)$ almost surely.

The proof of theorem 3.8 now follows by replacing all references to lemma 4.4 in lemmas 4.5, 4.7, and 4.8 by lemma 4.18.

After proving the asymptotic properties of the three adaptive estimators and showing that they have optimal asymptotic behaviors among the class of trimmed means, their behaviors in small samples will now be investigated through simulation in chapter 5.
Chapter 5

Small Sample Simulations

After investigating the asymptotic properties of some adaptive trimmed means, this chapter will concentrate on their finite sample behaviors. In the next section, the results of a Monte Carlo simulation are reported. Various adaptive trimmed means, distributions, and sample sizes are considered. Section 5.3 will be devoted to the results of a small Monte Carlo simulation of the coverage probabilities of some of the different confidence intervals proposed in chapter 3.

5.1. Finite Sample Variances

In this section we investigate the finite sample behavior of different adaptive trimmed means. Since the finite sample distribution of the \( \alpha \)-trimmed mean is not mathematically tractable, it is not surprising that the finite sample distribution of an adaptive trimmed mean is even less tractable. In this instance, as for most estimators (except those that are "distribution free"), the finite sample distribution is unknown. For a few estimators, such as the mean, even though one doesn’t know the finite sample distribution (unless the parent distribution is closed under convolution), certain characteristics of it, such as its moments, are known. But in most cases, not even the finite sample variance is known.

So how can we compare the finite sample behavior of the different estimators introduced in chapter 3? We have to resort to a Monte Carlo simulation. Obviously, this is not totally satisfactory. Since only a small number of distributions can reasonably be considered, how can we be assured that they are good representatives of families of distribution functions, not to mention data sets, usually encountered? There is obviously no good answer to that, but we can at least draw upon the vast experience of previous researchers
in Monte Carlo simulations of robust estimators. Foremost is the Princeton Robustness Study (Andrews et al., 1972) which assessed the performance of some 65 estimators in 32 different sampling situations including samples of size 5, 10, 20, and 40. Since then, the finite sample properties of most new (robust) estimators have often been analysed through simulations based on a subset of those 32 different sampling distributions. We shall do the same thing.

Let's first describe the different estimators that will be considered. They all share the same structure as outlined in section 2.3. Let $S_n(\alpha, F)$ be a functional. Then define

$$
\alpha_n(\hat{F}_n) = \arg \min_{\alpha \in A_n} S_n(\alpha, \hat{F}_n),
$$

where $\hat{F}_n$ is the empirical distribution function of a sample of size $n$ from $F$, and $A_n$ is defined in (3.2). Since $n$ will be considered fixed, we can write

$$
A_n = \{ \alpha : \alpha \in [\alpha_0, \alpha_1] \text{ and } n\alpha \in \mathbb{N} \}.
$$

In other words, $\alpha_n(\hat{F}_n)$ minimizes the value of $S_n(\alpha, \hat{F}_n)$ among all $\alpha$'s in an interval $[\alpha_0, \alpha_1]$ such that $n\alpha$ is an integer, so that only trimmed means trimming an integer number of observations are considered. The estimators are then $T(\alpha_n(\hat{F}_n), \hat{F}_n)$.

Five different functionals $S_n(\alpha, F)$ will be considered. The first three have been introduced in chapter 3. They are $S_n^{\text{Jac}}(\alpha, F)$, $S_n^{\text{Var}}(\alpha, F)$, and $S_n^{\text{BVAR}}(\alpha, F)$ leading to Jaekel's estimate (JAE), and the bootstrap adaptive trimmed means based on the variance (BVAR), and the interquartile range (BIQR) of the finite sample distribution of the $\alpha$-trimmed mean, respectively. All three random trimming proportions are selected through bootstrap estimates. The bootstrap estimates in JAE are expressible in closed form which is not the case for the other two. Therefore, one uses a Monte Carlo approximation to the bootstrap estimates in BVAR and BIQR. This approximation will be based on 100 bootstrap samples.

Two more estimators have been added as possible competitors to using the bootstrap in the selection of the trimming proportion. The first such method is cross-validation. The
idea is simple: split the data set in two subsets (not necessarily of the same size), then fit a
model on one subset and test the fit on the other subset. If one has more than one estimator
(model), then selecting that one which optimizes the fit seems like a reasonable method.
Indeed cross-validation is used to select the smoothing parameter in spline functions (e.g.
see Wahba and Wold (1975) and Craven and Wahba (1979)), and in density estimation
(e.g. see Rudemo (1982) and Bowman (1984)). In fact, in the case of kernel density
estimation, cross-validation is asymptotically optimal in a certain sense as demonstrated
in Hall and Marron (1987). On the other hand, Stone (1977) discusses some asymptotics
for and against cross-validation. For a general exposition to cross-validation, see Stone
(1974).

Let's now describe the cross-validation estimate. So as to simplify notation, I shall
write \( S_n(\alpha, \hat{F}_n) \) in terms of the \( X \)'s rather than as a function of \( \hat{F}_n \). By the way, in this
case, \( S_n(\alpha, F) \) is only defined for empirical distribution functions \( \hat{F}_n \), not for general \( F \)'s.
Let

\[
S_n(\alpha, \hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} \left[ X_i - T(\alpha, \hat{F}_n^{(-i)}) \right]^2,
\]

(5.2)

where \( \hat{F}_n^{(-i)} \) is the empirical distribution function of the \( X \)'s with the \( i \)th data point
removed. Note that \( S_n(\alpha, \hat{F}_n) \) is like an estimate of the mean squared error of prediction
where \( X_i \) is predicted by the \( \alpha \)-trimmed mean of the remaining data points after having
removed \( X_i \). To keep things simple, we did not quite use that definition of \( S_n(\alpha, \hat{F}_n) \).
Let \( \alpha = k/n \) where \( k \) is an integer. Instead of using a trimmed mean with trimming
proportion \( \alpha \) in the definition of \( S_n \), the trimming proportion \( k/(n - 1) \) was used. In this
way exactly \( k \) observations were trimmed on either side. This slight modification should
have very little effect on the results of the simulation. The adaptive trimmed mean based
on a cross-validation estimate of variance will be denoted CV.

The last adaptive estimator that will be considered is based on the jackknife. For a
review of the subject consult Miller (1974). The jackknife is a natural estimate to include
in this study as it is, in a certain way, the ancestor of the bootstrap. Using the notation
of Miller (1974), let \( \hat{\theta}_{-i} \) be \( T(\alpha, \hat{F}_n^{(-i)}) \), and let \( \hat{\theta}_i = n\hat{\theta} - (n - 1)\hat{\theta}_{-i} \), where \( \hat{\theta} = T(\alpha, \hat{F}_n) \).
Then let

$$S_n(\alpha, \hat{F}_n) = \text{Variance}(\hat{\theta}_i).$$

Note that $S_n(\alpha, \hat{F}_n)$ is also equal to the variance of the $\hat{\theta}_{-i}$'s. Again as was the case with cross-validation, we used trimming proportion $\alpha = k/(n - 1)$ instead of $\alpha = k/n$ in the definition of $\hat{\theta}_{-i}$. The adaptive trimmed mean based on a jackknife estimate of variance will be denoted JACK.

To complete the definition of these adaptive estimators, one has to specify the interval $[\alpha_0, \alpha_1]$ in the definition of $A_n$ in (5.1). We shall use three different intervals: [0, .25], [0, .50], and [.10, .40]. The fifteen different estimators will be denoted by JAE 0-25, JAE 0-50, JAE 10-40, ..., JACK 0-25, JACK 0-50, and JACK 10-40. Note that only the estimators 10-40 satisfy the conditions of chapter 3. Recall that assumption A of chapter 3 requires $\alpha_0$ to be strictly greater than 0. Neither of the estimators 0-25 or 0-50 satisfy this condition. Moreover, the theorems of chapter 3 hold only when $\alpha_1$ converges to .50 at a certain rate. Being only interested in one fixed sample size at a time, that criterion is not violated. But still, that requirement warns us that the estimators 0-50 might not work very well.

Let's reiterate that we are interested in the finite sample properties of the adaptive trimmed means. Therefore, the fact that some conditions of theorems about their asymptotic distributions are not met should not stop us from considering such estimators, especially for small sample sizes such as will be considered here. The estimator JAE 0-25 was investigated by the Princeton Robustness Study which explains the inclusion of the 0-25 estimators. Since the estimators 0-50 consider all trimmed means from the mean to the median that trim an integer number of observations, it is also a natural class of estimators. Because they satisfy the conditions set forth in chapter 3, the estimators 10-40 have been included as a compromise.

Finally, we also include all trimmed means with fixed trimming proportion $k/n$ for $k = 0, 1, \ldots, n - 1$. They will be denoted by T0, T5, and so on, where the number indicates the percentage of trimming. Note that all of the estimators introduced so far are invariant
under location and scale transformations.

Let's now introduce the four different sampling situations. Three of them consist of independently and identically distributed random variables from the standard normal, slash, and double exponential distributions. The other sampling situation is called the one-wild and is made up of \( n \) independent random variables with \( n - 1 \) of them from the standard normal distribution while the other one, the wild one, is from a normal distribution with variance equal to 100.

The normal distribution, being the assumed distribution in so many data analyses, is a natural choice for any simulation. Its tails decrease at a rate \( \exp(-x^2) \). As is well known, the mean is the optimal estimate of the location of the normal distribution for almost all conceivable criterions. In particular, the mean has the smallest variance among all trimmed means for every sample sizes. Note once again that, according to the asymptotic theory developed in chapter 3, the mean should not be included as one of the possible candidates of our adaptive estimators. But by letting \( \alpha_0 \) be arbitrarily close to 0, one could get an asymptotic variance arbitrarily close to that of the mean.

The slash distribution, defined to be the ratio of a standard normal to an independent uniform on \((0, 1)\), has been used in numerous simulation studies as a very long-tailed distribution. It doesn't have any moments and its tails decrease at the very slow rate of \( 1/x \). Its (disproportionate?) emphasis in simulation studies is probably due to the fact that it is expressible as the ratio of a normal to an independent variable. This allows the use of the normal over independent swindle (see e.g., Andrews et al. (1972)) which drastically improves the accuracy of the finite sample estimates of the variance of the various estimators. This advantage is unimportant to us as we will use a different swindle. More on that later.

The double exponential distribution is somewhat of a compromise between the normal and the slash. It is longer-tailed than the Gaussian, but shorter-tailed than the slash; its tails decrease at a rate of \( \exp(-x) \). Unlike the slash, it has moments of all orders, but it does not satisfy Winsor's principle as quoted in Tukey (1960), that is, it is not "normal
in the middle”.

Nevertheless, there is another important reason to include the double exponential. As is well known, the median is the maximum likelihood estimator for the double exponential distribution. Therefore, the median is asymptotically the best trimmed mean for that distribution. This contradicts assumption (A.5) of chapter 3. On the other hand, the median does not minimize the finite sample variance among the trimmed means. See figure 3.2. Even though the double exponential’s tails might not be as heavy as some other distributions, its presence in this simulation, along with the normal, will demonstrate the flexibility (or lack of) of the adaptive trimmed means.

The one-wild is supposed to be a good approximation to real life samples which often include one outlier. This is similar to the contaminated normals mentioned in chapter 2. There is some evidence that the one-wild sampling situation is more challenging than the contaminated normals i.i.d. sampling scheme. The interested reader should consult chapter 10 of “Understanding Robust and Exploratory Data Analysis” (URED) edited by Hoaglin, Mosteller, and Tukey (1983). In fact that chapter contains a lot of useful information about these and other distributions, along with exact finite sample results about a variety of L-estimators such as the trimmed means which will subsequently be used in our analysis.

Note that the four sampling situations are symmetric. As was discussed in chapter 3, non-symmetric distributions cause many problems such as the definition of the location of the distribution and the quantity being estimated by an adaptive estimator. Of course, to be realistic one should include such distributions. But this is a much harder problem that would require a very serious study of its own and is beyond what can be done here. Thus, for better or for worse, we stick to symmetric sampling situations.

The sample sizes considered in this simulation are 10, 20, and 40 for all fifteen estimators and four sampling situations.

The simulations were performed on a SUN 3/50 workstation. For each distribution and sample size, 10,000 samples were generated. To increase accuracy, the score swindle
of Johnstone and Velleman (1985) was used. It seems to be superior to the Gaussian over independent swindle for the distributions considered, and handles the one-wild sampling situation which the Gaussian over independent cannot. The Gaussian and double exponential random variables were generated by the ziggurat method of Marsaglia and Tsang (1984). Uniform random variables were generated by the IMSL function GGUBFS which is a linear congruential generator.

Let's now consider the question of performance. Because all of the estimators being under consideration are unbiased, it seems reasonable to use the variance as our criterion for judging their performance. In order to compare among sampling situations, we must standardize those variances in some way. Hence, we shall compute efficiencies, that is, compute the ratio of the variance of a reference estimator for the particular sampling situation to the variance of the estimator under study. We must therefore choose a reference estimator for each of the four sampling situations. To do that, let's recall that one of the two goals that were set in chapter 3 was to do as well as possible in small samples among the class of trimmed means. Therefore, it seems natural to use the best trimmed mean for the sampling situation and sample size under study as the reference estimator. Even though the adaptive estimators under study only consider trimming proportions of the form \( \alpha = k/n \) where \( n \) is the sample size and \( k \) is an integer, the reference estimator will be chosen among all trimmed means, i.e., all trimming proportions between 0 and .50 are considered. So formally, we define the efficiency of the estimator \( E \) at sampling situation \( F \), denoted by \( \text{eff}_F(E) \), by

\[
\text{eff}_F(E) = \frac{\text{Variance}_F(T_{\text{min}}(F))}{\text{Variance}_F(E)},
\]

where \( T_{\text{min}}(F) \) is the trimmed mean that minimizes the finite sample variance for the sampling situation \( F \) and sample size \( n \). To simplify the notation, the dependence on the sample size has been left out. In most other simulation studies, the reference estimator is the one which minimizes the variance among all estimators. But since our estimators select a member from the class of trimmed means, it would be unfair to use estimators outside of that class as our reference estimator given the fact that we are only interested
in comparing the adaptive trimmed means among themselves.

Tables 6.1 through 6.3 contain the efficiencies of the 15 adaptive trimmed means and of all trimmed means with fixed trimming proportion $k/n$ for $k = 0, 1, \ldots, (n-1)$ where $n$ is the sample size. The first table contains the results for sample size 10 whereas the last two contain the results for sample sizes 20 and 40, respectively. The variances of the reference estimators for sample sizes 10 and 20 were obtained from chapter 10 of UREDA. Since a trimmed mean is a linear combination of order statistics, one could in principle compute the exact variance by knowing the covariance matrix of the order statistics. Tables of order statistics for the four sampling situations under study and samples of size 10 and 20 exist and are referenced in UREDA. Therefore, the variance of the reference estimators, for these sample sizes, are exact (or as exact as numerical integration can be). On the other hand, such tables do not exist for samples of size 40. Hence, the reference estimators for table 6.3 were chosen to be the trimmed mean with fixed trimming proportion $k/40$ where $k = 0, 1, \ldots, 39$, such that the Monte Carlo estimate of variance is minimized. Notice the slight difference in the definition of the reference estimator for 40 observations.

For all sample sizes, the reference estimator for the normal distribution is the mean. In the case of the one-wild, the reference estimators are the 16%, 9%, and 5% trimmed means for samples of size 10, 20, and 40 respectively. The reference estimators for the slash are the 38%, 34%, and 32.5% trimmed means for the respective sample sizes. Finally, the 34%, 37%, and 40% trimmed means are the reference estimators for the double exponential.

As we noted before, the different estimators should be expected to perform differently at the different sampling situations depending, at the very least, on the range of trimming proportions considered. Figures 3.1 and 3.2 show the finite sample variance of the $\alpha$-trimmed means for samples of size 10, and 20, along with the asymptotic variance for the Gaussian and the double exponential distributions, respectively. Figure 5.1 shows the asymptotic variance for the slash distribution. These figures should clearly illustrate that point.
We are looking for an estimator that will do well for different sampling situations. This suggests optimizing the efficiency over different sampling situations. The concept which has been used most often seems to be that of polyefficiency, which apparently first appeared in Tukey (1979). This concept is a minimax approach. The polyefficiency of an estimator over a set of sampling situations is the minimum efficiency over the set of sampling situations. The idea is then to maximize the polyefficiency over the different estimators. One of the most common subset of sampling situations contains the normal, one-wild, and slash sampling situations. See chapter 11 of UREDA for instance. Then, define triefficiency 1 as follows:

\[ \text{trierf } 1(E) = \min \{ \text{eff}_{\text{normal}}(E), \text{eff}_{\text{one-wild}}(E), \text{eff}_{\text{slash}}(E) \}. \]

Some people don’t like putting too much emphasis on a distribution which is so heavy-tailed that it doesn’t even have moments. Hence a second subset will be considered where
the slash will be replaced by the double exponential. Let triefficiency 2 be

$$\text{trifeff } 2(E) = \min\{\text{eff}_{\text{normal}}(E), \text{eff}_{\text{one-wild}}(E), \text{eff}_{\text{double exponential}}(E)\}.$$ 

Of course some people will prefer the first subset to the second because, as was mentioned before, the double exponential fails to satisfy Winsor's principle. For the most pessimistic, we define the following quadefficiency,

$$\text{quadeff}(E) = \min\{\text{eff}_{\text{normal}}(E), \text{eff}_{\text{one-wild}}(E), \text{eff}_{\text{slash}}(E), \text{eff}_{\text{double exponential}}(E)\}.$$ 

A good estimator according to the quadefficiency criterion might be expected to do quite well in most situations as it will have been shown to do well for the normal distribution, which is the usual ideal model, the one-wild, which includes one outlier, the slash, which doesn't have any moments, and the double exponential, which requires a large trimming proportion unlike the normal. It should be noted that some people might object to a minimax criterion to select among the different estimators. Tukey (1979) has looked at some other criterions.

Tables 5.4 through 5.6 contain the three polyefficiencies corresponding to the efficiencies found in tables 5.1 through 5.3, respectively. We now have all the information necessary to compare the different adaptive estimators. We will start by looking at the truly bad estimators. Having then reduced the field of possible candidates, we will take a closer look at the remaining ones.

First, consider the cross-validation estimators CV. Let's first concentrate on the efficiencies at the slash. The efficiencies for CV 0-25 are 25.8, 16.0, and 8.1 for 10, 20, and 40 observations respectively. Likewise, they are 27.8, 16.7, and 8.3 for CV 0-50 and 49.5, 57.7, and 65.4 for CV 10-40. These are very low efficiencies. Even though things look better for CV 10-40, it is deceiving. The comparison must be made with respect to the trimmed means with trimming proportions in the same range. For instance, the efficiencies of the trimmed means with \(\alpha \in [0.1, 0.4]\) for a sample of size 40 vary from 56.1 to 100. The 10% and 12.5% trimmed means have an efficiency of 56.1 and 67.8. Contrast that with 65.4 for CV 10-40. In fact, all CV estimators, except CV 0-25 for 10 and 40 observa-
tions, are dominated, i.e., there is another estimator (for the same sample size) which has a larger efficiency at each of the four sampling situations. Moreover, they are also dominated from a polyefficiency point of view by one among JAE 0-25, and any of the BVAR and BIQR. They are also dominated when we disregard the slash distribution, as is done in triefficiency 2. It is easy to understand the extremely poor behavior of cross-validation at the slash. Expanding the square in (5.2), we see that the cross-validation estimate of variance is 
\[ \frac{1}{n} \left[ \sum X_i^2 - 2 \sum X_i T(\alpha, \hat{\delta}_n^{(-i)}) + \sum T^2(\alpha, \hat{\delta}_n^{(-i)}) \right]. \]
Even if the last term has a finite expectation, the first one does not. Therefore, this estimator has an infinite expectation for all values of \( \alpha \) whenever \( X \) doesn’t have a second moment. This leads to a very bad adaptive choice of \( \alpha \). Thus, the fact that the selection statistic can have an infinite expectation even when the given trimmed mean doesn’t should be sufficient to disqualify the cross-validation trimmed means. Consequently, they will no longer be considered.

Let’s now consider the jackknife estimators JACK. The first thing to notice is that they are very close to Jaeckel’s estimators JAE. The median absolute difference between the efficiencies of JAE and JACK for all sample sizes and all ranges of trimming proportions considered is 1.75. Is that surprising? Not really. Efron (1979, 1982) has shown that the jackknife estimate of variance of the functional \( \theta(\hat{F}_n) \) is the bootstrap estimate of variance of (approximately) the first order approximation of the functional \( \theta(\cdot) \) at \( \hat{F}_n \). The interested reader should consult those references, but here is the idea. Often, as is the case with the trimmed mean functional, one can expand the functional \( \theta(\hat{F}_n) \) as follows,

\[ \theta(\hat{F}_n) = \theta(F) + (1/n) \sum l_F(X_i) + Rem(\hat{F}_n, F). \]

When the remainder can be neglected, the asymptotic distribution of \( \theta(\hat{F}_n) \) is the same as that of \( (1/n) \sum l_F(X_i) \). Hence the asymptotic variance of \( \theta(\hat{F}_n) \) is exactly the variance of the random variable \( l_F(X) \). The infinitesimal jackknife estimate of variance of Jaeckel (1972) is \( (1/n^2) \sum l_F^2(X_i) \), i.e., it is a (bootstrap) estimate of the asymptotic variance of \( \theta(\hat{F}_n) \). The ordinary jackknife replaces the derivative \( l_F(X_i) \) by a finite difference, so that the ordinary jackknife can be viewed as an approximation to the infinitesimal jackknife. Hence, the ordinary jackknife estimate of variance of the \( \alpha \)-trimmed mean is an
approximation to Jaeckel's estimate of the asymptotic variance of the $\alpha$-trimmed mean. It is therefore not surprising that the two classes of estimators give very similar results. Also note that the polyefficiencies of the JACK estimators are dominated by either JAE 0-25 or any of the estimators BVAR or BIQR. Therefore the jackknife estimates JACK will not be considered anymore.

Let us turn to Jaeckel’s estimates JAE. First, note that JAE 0-25 behaves very differently from the other two. It does pretty well for all sample sizes, especially compared with the other 0-25 estimators. In general, it does not do as well as the other 0-25 estimators at the Gaussian and one-wild situations, but it usually does better at the slash and double exponential. This illustrates the fact that the estimators JAE seem to overemphasize the large trimming proportions. Whereas this hurts for the normal and the one-wild, it does not hurt enough to cancel the benefits obtained at the slash and the double exponential because the trimming proportion is at most 25%. Recall that trimming proportions in the 30%-40% range minimizes the finite sample variance of the $\alpha$-trimmed means for the slash and the double exponential distributions. While such an emphasis on larger trimming proportion is beneficial for JAE 0-25, it is disastrous for JAE 0-50, and JAE 10-40. Note, for instance, that the efficiency at the Gaussian for a sample of size 40 is 68.7 and 78.9 for JAE 0-50 and JAE 10-40, respectively. This should be contrasted with a range of efficiencies of 65.9 to 100.0 for $\alpha \in [0, .5]$ and a range of 72.0 to 94.5 for $\alpha \in [.1, .4]$. Looking at the polyefficiencies, we note that JAE 0-50 and JAE 10-40 are dominated by both bootstrap adaptive trimmed means at the corresponding range of trimming proportions for all sample sizes except for JAE 0-50 at samples of size 10 which is only dominated by BIQR 0-50.

To understand better what is happening, one has to look at the distribution of the estimates of the asymptotic variance. Note that this distribution is estimated by the empirical distribution function of the 10,000 simulated samples. For a sample of size 10 from the normal distribution, the means of the estimates of the asymptotic variance of the $\alpha$-trimmed mean with trimming proportions 0%, 10%, through 40% are 0.90, 0.93, 0.96,
0.96, and 0.70, while their medians are 0.83, 0.83, 0.80, 0.67, and 0.21. These numbers should be compared to 1.00, 1.06, 1.14, 1.25, and 1.39. The estimates of the asymptotic variance are clearly not very good for $\alpha$ beyond 20%, at least for samples of size 10. For samples of size 20, the story is much the same with the means of the estimates of the asymptotic variance increasing until $\alpha = .35$ where they then decrease, while the medians stop increasing at $\alpha = .25$. The same qualitative remarks about severe downward bias for large trimming proportions hold for the other distributions. Hence, $S_{Jae}^{\alpha}(\alpha, \hat{F}_n)$ can only be trusted on the range $\alpha \in [0, .25]$. Outside of that range, it is highly unreliable. Because of its poor performance at the normal distribution, both JAE 0-50 and JAE 10-40 are dominated in the polyefficiency tables. They will therefore be ignored for the rest of the discussion.

We are now left with JAE 0-25, and the 6 bootstrap adaptive trimmed means BVAR and BIQR. Having reduced the field of contestants from 15 down to 7, it is time to plot the results of tables 5.1 through 5.3 for this subset of estimators. Figure 5.2 contains plots of the efficiency of the seven estimators against sample size for the four sampling situations. The legend identifies the different estimators. Notice that full lines are used for the 0-25 estimators, whereas short broken lines and long broken lines are used for the 0-50 and 10-40 estimators, respectively.

By looking at the plot for the normal distribution, we notice that the two best estimators are 0-25 estimators, namely BIQR and BVAR, with BVAR 0-25 dominating all of the others. As expected, the 0-25 do better than the other ones although JAE 0-25 does not do as well as the other two. This is not surprising in view of the poor performance of $S_{Jae}^{\alpha}(\alpha, \hat{F}_n)$ as an estimator of the asymptotic variance. The lowest efficiency of the trimmed means for the normal is 72.3, 68.0, and 65.9 for 10, 20, and 40 observations. In general, all seven estimators are reasonable, with BVAR 0-25 being extremely good.

The plot for the one-wild is qualitatively similar to the previous one. The 0-25 estimators clearly outperform all of the other ones. This is not surprising as the one-wild is not a very demanding situation for the class of trimmed means: if there is only one
Figure 5.2: Efficiencies of the best adaptive trimmed means

Normal

One-wild

Slash

Double Exponential

1: Jae 0-25, 2: Bvar 0-25, 3: Bigr 0-25, 4: Bvar 0-50
5: Bigr 0-50, 6: Bvar 10-40, 7: Bigr 10-40
Section 5.1: Finite Sample Variances

outlier, any reasonable adaptive choice of the trimming proportion should favor trimming at least one observation on each side. Moreover, restricting the possible trimming to at most 25% should guarantee that such estimators work well. The lowest efficiency for the one-wild is 84.3, 75.7, and 70.8 for samples of size 10, 20, and 40. Most of them do reasonably well, with all three 0-25 doing very well.

As expected, the plot for the slash is radically different. The 10-40 are the best, while the 0-50 outperform the 0-25. In fact, the 0-25 are so bad for 10 observations that their efficiencies fall way off the plot; even BVAR 0-50 for 10 observations does not appear. Since the first and last two order statistics of a sample from the slash do not have a second moment, any trimmed mean which includes either of these order statistics will have an infinite variance. Hence for 10 observations, both the mean and the 10% trimmed mean have an infinite variance. Note that the estimated efficiency of the 10%, 5%, and 2.5% trimmed means for samples of size 10, 20, and 40 respectively are 18.7, 14.0, and 7.0, whereas the true efficiency is 0 for all of them. This shows that despite the use of a swindle and 10,000 simulated samples, the slash distribution is difficult to simulate.

Consider for instance samples of size 10. The estimated variances of the 0%, 10% through 40% trimmed means (multiplied by the sample size) are 172,227, 37.5, 9.3, 7.2, and 7.1. Note that the true value for the first two is infinity. The means of the estimates of the asymptotic variance $S^{Jac}(\alpha, \hat{F}_n)$ were 155,119, 46.4, 8.7, 6.0, and 3.6 while the medians were 20.4, 6.5, 4.5, 3.3, and 1.0. These estimates are not too bad except when the trimming proportion is too large. The interesting point is that even though there is quite a large spread between the means, the spread is much tighter for the medians. This makes it much harder to differentiate among the different trimmed means. Let's now look at the bootstrap estimate of variance for the same situation. The means are 160,308, 123,414, 65,021, 33,354, and 11,596, while the medians are 20.0, 19.9, 15.6, 11.5, and 9.7. These numbers clearly illustrate the problems that these estimators have with the slash distribution for samples of size 10.

Looking at the slash with a sample of size 20, some improvement is found. The
bootstrap estimates BVAR 0-25 and BIQR 0-25 are still pretty bad, but the others are much better. Throughout, we noticed that the bootstrap estimates of the finite and asymptotic variance were best for trimming proportions away from 0 and 0.5. This is of course consistent with the theory of chapters 3 and 4 which required that the largest trimming proportion considered tend to 0.5 at a specific rate and that the smallest be fixed and strictly positive in order for the estimates to converge uniformly. In remark 4.2, we suggested that it might be possible to allow the smallest trimming proportion to converge to 0 by adding extra conditions, but once again, only at a certain rate. This last provision seems to be justified empirically, not only at the slash, but also at the other distributions.

As the sample size increases further to 40, all estimators do a good job of adapting themselves. The bootstrap estimates of finite and asymptotic variance are much better, especially in the center. Near 0.5, as usual, the estimates of the asymptotic variance show their usual downward bias. Such a phenomenon is not as pronounced in the estimates of the finite variance.

Let’s now turn to the double exponential distribution. Qualitatively, the plots for the double exponential and the slash are very similar, except for samples of size 10. That is, the estimators 10-40 and 0-50 are better than the 0-25. Of course, there is a major difference in that even though the worst trimmed mean in both cases is the mean, its efficiency at the slash is 0 whereas it is 70.2, 63.8, and 59.7 for the double exponential at 10, 20, and 40 observations respectively. Hence the adaptive trimmed means cannot do as badly as in the case of the slash. Notice that BIQR 10-40 dominates the six other estimators.

Figure 5.2, although useful, is not sufficient to help us in the selection of an adaptive trimmed mean. Let’s look at tables 5.4 through 5.6 to see how the polyefficiencies can help choose among those seven adaptive estimators. According to the criterion triefficiency 1, BIQR 0-50 is the best for samples of size 10, whereas BVAR 0-50 and BVAR 0-25 are best for samples of size 20 and 40, respectively. For triefficiency 2 (which replaces the slash by the double exponential), BVAR 10-40, JAE 0-25, and BVAR 0-50 are the best
for the usual sample sizes. Finally, BIQR 0-50, BVAR 10-40, and BVAR 0-50 maximizes quadefficiency at the respective sample sizes. Notice that each of these "best" adaptive estimators outperform the respective "best" non-adaptive trimmed means, the differences ranging from 1.1 to 7.4.

That still leaves us short of finding the adaptive trimmed mean. If one takes an overall minimax approach of maximizing the worst possible efficiency over the twelve possible combinations of sampling situation and sample size, then the winner is BIQR 10-40 followed by BIQR 0-50 with minimum efficiencies of 83.2 and 82.2 respectively. The corresponding winner for the non-adaptive trimmed means is T30 with a minimum efficiency of 79.8. The estimator BVAR 10-40 with a minimum efficiency of 81.7 also does better than the trimmed mean T30. Note that this estimator does better than BIQR 10-40 for the normal and the one-wild while doing slightly worse at the slash and double exponential. It only does much worse at the slash for a sample of size 10 where it achieves its minimum efficiency. Hence, these arguments, along with the fact that it satisfies the conditions of chapter 3, make BVAR 10-40 a very good adaptive trimmed mean. Of course, other optimizing criterions would lead to other recommended choices. We leave it to the reader to select his/her preferred adaptive trimmed mean.

In the next section, confidence intervals based on adaptive trimmed means will be investigated.

5.2. Bootstrap Confidence Intervals

In chapter 3, several confidence intervals for the location of a distribution based on adaptive trimmed means were presented. They fell in two categories: those based on the asymptotic distribution and those based on the bootstrap distribution. The small sample properties of such confidence intervals were investigated in a small Monte Carlo study whose results will now be reported.

Four different confidence intervals were considered. First is the standard (asymptotic) confidence interval based on JAE 0-25. This interval uses an estimate of the asymptotic
variance of the trimmed mean at the adaptively chosen trimming proportion. As we have seen, those estimates are not very reliable for $\alpha$ much larger than 0.25, which is why we have chosen the range 0-25. The interval is given in formula (3.6). The second interval is the bootstrap confidence interval based on JAE 0-25 of (3.7). This interval comes from approximating the finite sample distribution of JAE 0-25 by its bootstrap distribution rather than its asymptotic distribution.

The last two intervals are based on BVAR 0-25. First is the bootstrap confidence interval given by (3.12). Remember that such a confidence interval involves a double bootstrap, as one bootstrap selects the trimming proportion to use while the second one gives the bootstrap distribution of the adaptive trimmed mean. Finally, the bootstrap-t based on BVAR 0-25 was also included to try to see if the second-order correctness of the bootstrap-t in some other contexts would empirically remain valid in this adaptive setting. This interval is given by (3.9) with all references to Jaeckel’s estimate being replaced by the bootstrap adaptive trimmed mean based on the finite sample variance. This interval involves computing an estimate of the asymptotic variance for each bootstrap sample of the outer loop. As in the previous interval, one level of bootstrapping (the inner loop) is used to select the trimming proportion of the bootstrap samples of the outer loop. Whereas the previous interval only requires the computation of BVAR 0-25 for that bootstrap sample, the bootstrap-t interval also requires the computation of an estimate of asymptotic variance of the trimmed mean with the adaptively chosen trimming proportion. Again, this is why we have used the range 0-25.

Because of the tremendous amount of computations involved, only 500 samples of size 10 were simulated at each of three different distributions. They are the standard normal, the slash, and the double exponential. The estimates BVAR 0-25 were based on 100 bootstrap replications. The bootstrap confidence intervals were computed from 1,000 bootstrap samples. Hence, for a single sample of 10 observations, the bootstrap confidence interval based on JAE 0-25 requires 1,000 bootstrap samples of size 10, whereas the last two confidence intervals require 100,000 bootstrap samples. This must be contrasted with
the standard interval based on JAE 0-25 which does not require any resampling.

For each of the simulated samples, upper 5% and 95% one-sided confidence intervals and the resulting 90% two-sided confidence interval were computed. In order to compare the four different methods, the estimated coverage probability of the three intervals were computed along with the mean and the standard deviation of the length of the 90% two-sided confidence intervals. Of course, whenever two different methods lead to intervals with very different coverage probabilities, comparing the length of the confidence intervals is not very useful, but in the other case, the length can be a useful measure.

The results of the simulations are found in table 5.7. Let's start with the Gaussian distribution. The coverage probabilities for the standard intervals based on JAE 0-25 are way off the mark. The actual distribution of the root $n^{1/2}[T_n^{Jae}(\hat{\theta}_n) - \theta]$ is much wider than its asymptotic approximation. Therefore, there is more than twice as much probability in each tail than claimed, leading to a coverage probability for the 90% two-sided confidence interval of 77.2%. Note that the standard deviation for these estimates is at most 2.2%. Such a result is highly undesirable, but not very surprising. As was noted in chapter 3, this interval fails to incorporate the adaptive nature of the estimator.

The results for the bootstrap intervals based either on JAE 0-25 or on BVAR 0-25 are better, although not great. The differences between the target coverage probabilities and their actual values for the one-sided intervals have almost been reduced in half from over 6% to about 3.5%. This leads to coverage probabilities for the two-sided 90% interval of 82.8% and 83.0%. This is still well under the target value.

On the other hand, the results for the bootstrap-t intervals based on BVAR 0-25 are outstanding. With estimated errors of 0.2%, and 0.6% for the one-sided intervals and 0.8% for the two-sided interval—keeping in mind that the standard deviation of these estimates is of the order of 2%—the bootstrap-t is not far from being exact. It is clear, in the case of the normal distribution, that if one is serious enough about the confidence coefficient of a confidence interval based on one of those adaptive trimmed means, then the bootstrap-t should be used in spite of its computational cost, i.e., 100,000 bootstrap
samples. Note that the mean lengths of the two-sided intervals are consistent with the estimated coverage probabilities; that is, given that the bootstrap-t is almost exact and that the other ones have lower coverage probabilities, they also have shorter intervals.

The results for the slash distribution are slightly different. First, the results for the standard interval are essentially as bad as they were in the case of the normal distribution. On the other hand, the bootstrap intervals of JAE 0-25 and BVAR 0-25 are much better. They are within one standard deviation of their target for the one-sided intervals and slightly over one standard deviation for the two-sided interval. Finally, the bootstrap-t doesn't do as well as it did for the Gaussian. While it has reasonable coverage probabilities, it seems to be outperformed by the other two bootstrap intervals.

The results about the length of the two-sided intervals are much harder to interpret. The standard interval is on average the smallest, followed by the bootstrap-t which is more than twice as large and has a much larger standard deviation. Then the ordinary bootstrap intervals follow with a mean 4 to 6 times larger than the bootstrap-t and with standard deviations much larger still. If these numbers are a good reflection of the reality (remember that the slash distribution does not have any moment and that only 500 samples of size 10 were simulated) then it certainly is tempting to use the bootstrap-t interval with a slightly lower coverage probability but a much narrower interval, on average, than the other two bootstrap intervals.

Finally, let's look at the double exponential distribution. Once again, the standard intervals are overly optimistic. The bootstrap intervals based on JAE 0-25 are better than those based on BVAR 0-25. It is not clear whether the difference is real, but if it is, it might be explained by the much higher efficiency of JAE 0-25 over BVAR 0-25 at the double exponential distribution for samples of size 10. Also, the bootstrap-t ended up between the other two bootstrap intervals. As for the lengths, even though the three bootstrap intervals have pretty much the same coverage probability for the double-sided interval, the bootstrap-t is significantly larger and more variable than the other two.

What should be concluded from these simulations? First, that the standard intervals
cannot be relied upon. They are much too optimistic. Second, that the bootstrap confidence intervals have reasonable coverage probabilities, especially in view of the fact that they were based on samples of size 10. Third, that the bootstrap-t intervals sometimes do much better than the ordinary bootstrap intervals, while they do worse for some other distributions. In view of the fact that the bootstrap interval based on JAE 0-25 only involves 1,000 bootstrap samples instead of the 100,000 of the other two, it seems clear that it should be the preferred choice.

In the next section, we conclude this chapter with a few remarks.

5.3. Remarks

Let’s conclude this chapter on finite sample properties of adaptive trimmed means with a few remarks.

**Remark 5.1:** The bootstrap estimators BVAR and BIQR were based on 100 bootstrap replications. Preliminary simulations were done using 1,000 bootstrap samples leading, most of the time, to slightly better results. But the gains/losses were usually of the order of 1%. Such gains do not seem to be worth a ten-fold increase in the computation time. Furthermore, it might even be possible to reduce the number of bootstrap samples even further. Recall that only the ordering of the variance (or interquartile range) estimates matter, not their precise values. Therefore, bootstrap sample sizes of 25 might be sufficient, although no simulations results are available to support that claim.

**Remark 5.2:** The efficiencies computed were ratios of variances. It might be interesting to use a more robust measure of spread (squared) to judge the performance of the different estimators. For instance, BVAR seemed to outperform BIQR most of the time, except for samples of size 10. Would that conclusion remain valid if the square of the interquartile range was used in the computation of efficiencies instead of the variance? If it does not, then which of the two criterions should be used?

**Remark 5.3:** As was noted in chapter 3, BIQR selects the trimming proportion that minimizes the length of a 50% bootstrap confidence interval. The theory would also
hold for any other confidence level. By looking at the distribution of the trimmed means under the various sampling plans, the upper and lower quartiles do not seem to vary as much as the upper and lower deciles, for instance. Therefore, minimizing the length of a central 80% confidence interval might lead to better small sample results than minimizing the interquartile range. But one must keep in mind that the 10th and 90th percentiles are more variable than the upper and lower quartiles, so that the (possibly) higher variability of the length of an 80% bootstrap confidence interval might actually lead to a worse choice of the trimming proportion, especially when those estimates come from only 100 bootstrap samples.

**Remark 5.4:** Even though the usual bootstrap and the bootstrap-t based on BVAR or BIQR are both double bootstrap procedures, the latter requires more storage in its algorithmic implementation. Remember from the discussion following the algorithm of the bootstrap confidence interval based on Jaeckel’s estimate in chapter 3, that it is only necessary to store the value of the adaptive trimmed mean of the bootstrap sample, rather than the difference between that value and the adaptive trimmed mean of the original sample. The bootstrap-t, on the other hand, requires the storage of the ratio of the difference of the adaptive trimmed means of the bootstrap sample and the original sample to an estimate of the asymptotic variance of the adaptive trimmed mean based on the bootstrap sample. Unfortunately, the adaptive trimmed mean of the original sample is not known until the outer bootstrap loop is completed. Therefore, both the adaptive trimmed mean of the bootstrap samples along with their estimate of asymptotic variance must be stored.

**Remark 5.5:** How do these estimators compare to other robust estimators of location? Rather well. Consider for instance triefficiency 1 for a sample of size 20. Using chapter 11 of UREDA, we find out that the efficiencies for the one-wild must be multiplied by 0.9217 when the reference estimator is replaced by the minimum variance estimator, defined to be the estimator with the smallest variance which is included in the Princeton
Robustness Study or its continuation. For the one-wild with sample size 20, this estimator is the biweight with tuning constant 8.8. The reference estimator for the slash with sample size 20 is the Pitman estimator, and the efficiencies must be multiplied by 0.9135. Then BVAR 10-40 has efficiencies of 87.8, 85.2, and 86.5 with a triefficiency 1 of 85.2. Compare with T25 which has efficiencies of 84.2, 85.2, 85.9, and triefficiency 1 of 84.2. Such results compare favorably with the biweight with tuning constant 6.0 which has a reported triefficiency 1 of 81.8, or with a Hampel with tuning constants of 1.7, 3.4, 8.5 which has a reported triefficiency 1 of 82.1. Yet, other estimators do better, such as a P-estimator of Johns (1979), and a biweight with tuning constant 6.4 which have triefficiencies 1 of 88 and 85, respectively. The adaptive trimmed mean BVAR 10-40 also falls short of the adaptive M-estimator of Bell (1980) which has a triefficiency 1 with respect to the aforementioned reference estimators of 90.0. Nevertheless, such results show that BVAR 10-40 not only does well when compared to the class of trimmed means, but it also does well among all robust estimators.
Table 5.1

Efficiencies of estimators based on a sample of size 10

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Normal</th>
<th>One-Wild</th>
<th>Slash</th>
<th>Double Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAE 0-25</td>
<td>92.0</td>
<td>98.6</td>
<td>64.6</td>
<td>89.5</td>
</tr>
<tr>
<td>BVAR 0-25</td>
<td>97.7</td>
<td>93.8</td>
<td>32.7</td>
<td>78.7</td>
</tr>
<tr>
<td>BIQR 0-25</td>
<td>95.5</td>
<td>96.3</td>
<td>57.1</td>
<td>83.1</td>
</tr>
<tr>
<td>CV 0-25</td>
<td>94.1</td>
<td>96.2</td>
<td>25.8</td>
<td>84.8</td>
</tr>
<tr>
<td>JACK 0-25</td>
<td>92.9</td>
<td>98.5</td>
<td>63.2</td>
<td>88.0</td>
</tr>
<tr>
<td>JAE 0-50</td>
<td>74.9</td>
<td>85.6</td>
<td>97.1</td>
<td>95.3</td>
</tr>
<tr>
<td>BVAR 0-50</td>
<td>95.7</td>
<td>88.5</td>
<td>59.3</td>
<td>81.7</td>
</tr>
<tr>
<td>BIQR 0-50</td>
<td>87.6</td>
<td>90.8</td>
<td>85.6</td>
<td>89.7</td>
</tr>
<tr>
<td>CV 0-50</td>
<td>84.9</td>
<td>86.8</td>
<td>27.8</td>
<td>89.1</td>
</tr>
<tr>
<td>JACK 0-50</td>
<td>78.6</td>
<td>87.8</td>
<td>94.3</td>
<td>94.2</td>
</tr>
<tr>
<td>JAE 10-40</td>
<td>74.3</td>
<td>85.5</td>
<td>97.6</td>
<td>95.9</td>
</tr>
<tr>
<td>BVAR 10-40</td>
<td>91.0</td>
<td>92.0</td>
<td>81.7</td>
<td>91.0</td>
</tr>
<tr>
<td>BIQR 10-40</td>
<td>84.0</td>
<td>92.2</td>
<td>91.8</td>
<td>94.5</td>
</tr>
<tr>
<td>CV 10-40</td>
<td>82.1</td>
<td>87.0</td>
<td>49.5</td>
<td>93.9</td>
</tr>
<tr>
<td>JACK 10-40</td>
<td>77.2</td>
<td>87.9</td>
<td>95.8</td>
<td>95.5</td>
</tr>
<tr>
<td>T0</td>
<td>100.0</td>
<td>12.9</td>
<td>0.0</td>
<td>69.9</td>
</tr>
<tr>
<td>T10</td>
<td>94.9</td>
<td>98.8</td>
<td>18.7</td>
<td>86.4</td>
</tr>
<tr>
<td>T20</td>
<td>88.3</td>
<td>98.7</td>
<td>74.9</td>
<td>95.5</td>
</tr>
<tr>
<td>T30</td>
<td>80.8</td>
<td>92.8</td>
<td>96.1</td>
<td>99.0</td>
</tr>
<tr>
<td>T40</td>
<td>72.2</td>
<td>84.0</td>
<td>98.4</td>
<td>95.7</td>
</tr>
</tbody>
</table>
Table 5.2

Efficiencies of estimators based on a sample of size 20

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Normal</th>
<th>One-Wild</th>
<th>Slash</th>
<th>Double Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAE 0-25</td>
<td>90.3</td>
<td>95.6</td>
<td>86.1</td>
<td>88.9</td>
</tr>
<tr>
<td>BVAR 0-25</td>
<td>96.6</td>
<td>96.0</td>
<td>76.3</td>
<td>81.3</td>
</tr>
<tr>
<td>BIQR 0-25</td>
<td>95.2</td>
<td>96.6</td>
<td>77.4</td>
<td>82.7</td>
</tr>
<tr>
<td>CV 0-25</td>
<td>95.8</td>
<td>94.7</td>
<td>16.0</td>
<td>78.2</td>
</tr>
<tr>
<td>JACK 0-25</td>
<td>93.3</td>
<td>97.4</td>
<td>74.7</td>
<td>84.2</td>
</tr>
<tr>
<td>JAE 0-50</td>
<td>70.5</td>
<td>77.6</td>
<td>95.0</td>
<td>95.1</td>
</tr>
<tr>
<td>BVAR 0-50</td>
<td>92.3</td>
<td>91.8</td>
<td>91.8</td>
<td>86.5</td>
</tr>
<tr>
<td>BIQR 0-50</td>
<td>83.8</td>
<td>87.1</td>
<td>94.7</td>
<td>91.4</td>
</tr>
<tr>
<td>CV 0-50</td>
<td>84.6</td>
<td>82.0</td>
<td>16.7</td>
<td>85.1</td>
</tr>
<tr>
<td>JACK 0-50</td>
<td>74.1</td>
<td>80.7</td>
<td>95.1</td>
<td>94.5</td>
</tr>
<tr>
<td>JAE 10-40</td>
<td>77.0</td>
<td>84.6</td>
<td>97.8</td>
<td>97.1</td>
</tr>
<tr>
<td>BVAR 10-40</td>
<td>87.8</td>
<td>92.4</td>
<td>94.7</td>
<td>92.3</td>
</tr>
<tr>
<td>BIQR 10-40</td>
<td>83.2</td>
<td>89.6</td>
<td>95.8</td>
<td>94.6</td>
</tr>
<tr>
<td>CV 10-40</td>
<td>86.7</td>
<td>89.6</td>
<td>57.7</td>
<td>90.9</td>
</tr>
<tr>
<td>JACK 10-40</td>
<td>79.3</td>
<td>86.6</td>
<td>96.6</td>
<td>96.0</td>
</tr>
<tr>
<td>T0</td>
<td>100.0</td>
<td>20.6</td>
<td>0.0</td>
<td>63.5</td>
</tr>
<tr>
<td>T5</td>
<td>97.7</td>
<td>98.4</td>
<td>14.0</td>
<td>74.5</td>
</tr>
<tr>
<td>T10</td>
<td>94.7</td>
<td>99.8</td>
<td>44.4</td>
<td>81.8</td>
</tr>
<tr>
<td>T15</td>
<td>91.4</td>
<td>98.2</td>
<td>68.3</td>
<td>87.5</td>
</tr>
<tr>
<td>T20</td>
<td>87.8</td>
<td>95.6</td>
<td>84.4</td>
<td>92.2</td>
</tr>
<tr>
<td>T25</td>
<td>84.2</td>
<td>92.4</td>
<td>94.0</td>
<td>96.0</td>
</tr>
<tr>
<td>T30</td>
<td>80.4</td>
<td>88.7</td>
<td>98.8</td>
<td>98.7</td>
</tr>
<tr>
<td>T35</td>
<td>76.4</td>
<td>84.5</td>
<td>100.1</td>
<td>99.7</td>
</tr>
<tr>
<td>T40</td>
<td>72.2</td>
<td>80.0</td>
<td>98.5</td>
<td>98.9</td>
</tr>
<tr>
<td>T45</td>
<td>67.6</td>
<td>75.0</td>
<td>94.5</td>
<td>95.6</td>
</tr>
</tbody>
</table>
Table 5.3
Efficiencies of estimators based on a sample of size 40

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Normal</th>
<th>One-Wild</th>
<th>Slash</th>
<th>Double</th>
<th>Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAE 0-25</td>
<td>91.7</td>
<td>94.9</td>
<td>92.0</td>
<td>86.2</td>
<td></td>
</tr>
<tr>
<td>BVAR 0-25</td>
<td>95.9</td>
<td>96.2</td>
<td>92.3</td>
<td>83.2</td>
<td></td>
</tr>
<tr>
<td>BIRQ 0-25</td>
<td>94.6</td>
<td>96.0</td>
<td>89.9</td>
<td>82.5</td>
<td></td>
</tr>
<tr>
<td>CV 0-25</td>
<td>96.7</td>
<td>93.7</td>
<td>8.1</td>
<td>74.3</td>
<td></td>
</tr>
<tr>
<td>JACK 0-25</td>
<td>93.4</td>
<td>96.3</td>
<td>88.4</td>
<td>84.0</td>
<td></td>
</tr>
</tbody>
</table>

| JAE 0-50  | 68.7   | 73.2     | 91.6  | 95.8   |             |
| BVAR 0-50 | 90.2   | 90.6     | 96.7  | 89.7   |             |
| BIRQ 0-50 | 82.2   | 84.8     | 94.9  | 92.7   |             |
| CV 0-50   | 84.5   | 81.3     | 8.3   | 82.0   |             |
| JACK 0-50 | 72.0   | 76.3     | 92.5  | 95.7   |             |

| JAE 10-40 | 78.9   | 83.8     | 97.3  | 96.1   |             |
| BVAR 10-40| 86.4   | 90.7     | 97.6  | 92.2   |             |
| BIRQ 10-40| 83.5   | 88.3     | 96.9  | 93.4   |             |
| CV 10-40  | 89.2   | 87.0     | 65.4  | 87.6   |             |
| JACK 10-40| 80.5   | 85.4     | 97.0  | 95.1   |             |

| T0        | 100.0  | 32.9     | 0.0   | 59.7   |             |
| T2.5      | 99.0   | 98.8     | 7.0   | 66.2   |             |
| T5.0      | 97.6   | 100.0    | 24.3  | 70.8   |             |
| T7.5      | 96.1   | 99.7     | 41.6  | 74.5   |             |
| T10.0     | 94.5   | 98.8     | 56.1  | 77.9   |             |
| T12.5     | 92.8   | 97.6     | 67.8  | 80.9   |             |
| T15.0     | 91.1   | 96.2     | 77.1  | 83.7   |             |
| T17.5     | 89.3   | 94.6     | 84.3  | 86.3   |             |
| T20.0     | 87.5   | 92.9     | 90.0  | 88.8   |             |
| T22.5     | 85.6   | 91.1     | 94.1  | 91.0   |             |
| T25.0     | 83.7   | 89.2     | 97.0  | 93.1   |             |
| T27.5     | 81.8   | 87.2     | 98.8  | 95.0   |             |
| T30.0     | 79.8   | 85.3     | 99.8  | 96.7   |             |
| T32.5     | 77.8   | 83.2     | 100.0 | 98.1   |             |
| T35.0     | 75.8   | 81.2     | 99.6  | 99.2   |             |
| T37.5     | 73.8   | 79.1     | 98.8  | 99.9   |             |
| T40.0     | 72.0   | 77.2     | 97.4  | 100.0  |             |
| T42.5     | 70.1   | 75.2     | 95.7  | 99.5   |             |
| T45.0     | 68.1   | 73.2     | 93.5  | 98.3   |             |
| T47.5     | 65.9   | 70.8     | 90.8  | 96.2   |             |
Table 5.4
Polyefficiencies of estimators based on a sample of size 10

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Triefficiency 1</th>
<th>Triefficiency 2</th>
<th>Quadefficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAE 0-25</td>
<td>64.6</td>
<td>89.5</td>
<td>64.6</td>
</tr>
<tr>
<td>BVAR 0-25</td>
<td>32.7</td>
<td>78.7</td>
<td>32.7</td>
</tr>
<tr>
<td>BIQR 0-25</td>
<td>57.1</td>
<td>83.1</td>
<td>57.1</td>
</tr>
<tr>
<td>CV 0-25</td>
<td>25.8</td>
<td>84.8</td>
<td>25.8</td>
</tr>
<tr>
<td>JACK 0-25</td>
<td>63.2</td>
<td>88.0</td>
<td>63.2</td>
</tr>
<tr>
<td>JAE 0-50</td>
<td>74.9</td>
<td>74.9</td>
<td>74.9</td>
</tr>
<tr>
<td>BVAR 0-50</td>
<td>59.3</td>
<td>81.7</td>
<td>59.3</td>
</tr>
<tr>
<td>BIQR 0-50</td>
<td>85.6</td>
<td>87.6</td>
<td>85.6</td>
</tr>
<tr>
<td>CV 0-50</td>
<td>27.8</td>
<td>84.9</td>
<td>27.8</td>
</tr>
<tr>
<td>JACK 0-50</td>
<td>78.6</td>
<td>78.6</td>
<td>78.6</td>
</tr>
<tr>
<td>JAE 10-40</td>
<td>74.3</td>
<td>74.3</td>
<td>74.3</td>
</tr>
<tr>
<td>BVAR 10-40</td>
<td>81.7</td>
<td>91.0</td>
<td>81.7</td>
</tr>
<tr>
<td>BIQR 10-40</td>
<td>84.0</td>
<td>84.0</td>
<td>84.0</td>
</tr>
<tr>
<td>CV 10-40</td>
<td>49.5</td>
<td>82.1</td>
<td>49.5</td>
</tr>
<tr>
<td>JACK 10-40</td>
<td>77.2</td>
<td>77.2</td>
<td>77.2</td>
</tr>
<tr>
<td>T0</td>
<td>0.0</td>
<td>12.9</td>
<td>0.0</td>
</tr>
<tr>
<td>T10</td>
<td>18.7</td>
<td>86.4</td>
<td>18.7</td>
</tr>
<tr>
<td>T20</td>
<td>74.9</td>
<td>88.3</td>
<td>74.9</td>
</tr>
<tr>
<td>T30</td>
<td>80.8</td>
<td>80.8</td>
<td>80.8</td>
</tr>
<tr>
<td>T40</td>
<td>72.2</td>
<td>72.2</td>
<td>72.2</td>
</tr>
</tbody>
</table>
Table 5.5
Polyefficiencies of estimators based on a sample of size 20

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Triefficiency 1</th>
<th>Triefficiency 2</th>
<th>Quadefficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAE 0-25</td>
<td>86.1</td>
<td>88.9</td>
<td>86.1</td>
</tr>
<tr>
<td>BVAR 0-25</td>
<td>76.3</td>
<td>81.3</td>
<td>76.3</td>
</tr>
<tr>
<td>BIQR 0-25</td>
<td>77.4</td>
<td>82.7</td>
<td>77.4</td>
</tr>
<tr>
<td>CV 0-25</td>
<td>16.0</td>
<td>78.2</td>
<td>16.0</td>
</tr>
<tr>
<td>JACK 0-25</td>
<td>74.7</td>
<td>84.2</td>
<td>74.7</td>
</tr>
<tr>
<td>JAE 0-50</td>
<td>70.5</td>
<td>70.5</td>
<td>70.5</td>
</tr>
<tr>
<td>BVAR 0-50</td>
<td>91.8</td>
<td>86.5</td>
<td>86.5</td>
</tr>
<tr>
<td>BIQR 0-50</td>
<td>83.8</td>
<td>83.8</td>
<td>83.8</td>
</tr>
<tr>
<td>CV 0-50</td>
<td>16.7</td>
<td>82.0</td>
<td>16.7</td>
</tr>
<tr>
<td>JACK 0-50</td>
<td>74.1</td>
<td>74.1</td>
<td>74.1</td>
</tr>
<tr>
<td>JAE 10-40</td>
<td>77.0</td>
<td>77.0</td>
<td>77.0</td>
</tr>
<tr>
<td>BVAR 10-40</td>
<td>87.8</td>
<td>87.8</td>
<td>87.8</td>
</tr>
<tr>
<td>BIQR 10-40</td>
<td>83.2</td>
<td>83.2</td>
<td>83.2</td>
</tr>
<tr>
<td>CV 10-40</td>
<td>57.7</td>
<td>86.7</td>
<td>57.7</td>
</tr>
<tr>
<td>JACK 10-40</td>
<td>79.3</td>
<td>79.3</td>
<td>79.3</td>
</tr>
<tr>
<td>T0</td>
<td>0.0</td>
<td>20.6</td>
<td>0.0</td>
</tr>
<tr>
<td>T5</td>
<td>14.0</td>
<td>74.5</td>
<td>14.0</td>
</tr>
<tr>
<td>T10</td>
<td>44.4</td>
<td>81.8</td>
<td>44.4</td>
</tr>
<tr>
<td>T15</td>
<td>68.3</td>
<td>87.5</td>
<td>68.3</td>
</tr>
<tr>
<td>T20</td>
<td>84.4</td>
<td>87.8</td>
<td>84.4</td>
</tr>
<tr>
<td>T25</td>
<td>84.2</td>
<td>84.2</td>
<td>84.2</td>
</tr>
<tr>
<td>T30</td>
<td>80.4</td>
<td>80.4</td>
<td>80.4</td>
</tr>
<tr>
<td>T35</td>
<td>76.4</td>
<td>76.4</td>
<td>76.4</td>
</tr>
<tr>
<td>T40</td>
<td>72.2</td>
<td>72.2</td>
<td>72.2</td>
</tr>
<tr>
<td>T45</td>
<td>67.6</td>
<td>67.6</td>
<td>67.6</td>
</tr>
</tbody>
</table>
### Table 5.6
Polyefficiencies of estimators based on a sample of size 40

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Triefficiency 1</th>
<th>Triefficiency 2</th>
<th>Quadefficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAE 0-25</td>
<td>91.7</td>
<td>86.2</td>
<td>86.2</td>
</tr>
<tr>
<td>BVAR 0-25</td>
<td>92.3</td>
<td>83.2</td>
<td>83.2</td>
</tr>
<tr>
<td>BIQR 0-25</td>
<td>89.9</td>
<td>82.5</td>
<td>82.5</td>
</tr>
<tr>
<td>CV 0-25</td>
<td>8.1</td>
<td>74.3</td>
<td>8.1</td>
</tr>
<tr>
<td>JACK 0-25</td>
<td>88.4</td>
<td>84.0</td>
<td>84.0</td>
</tr>
<tr>
<td>JAE 0-50</td>
<td>68.7</td>
<td>68.7</td>
<td>68.7</td>
</tr>
<tr>
<td>BVAR 0-50</td>
<td>90.2</td>
<td>89.7</td>
<td>89.7</td>
</tr>
<tr>
<td>BIQR 0-50</td>
<td>82.2</td>
<td>82.2</td>
<td>82.2</td>
</tr>
<tr>
<td>CV 0-50</td>
<td>8.3</td>
<td>81.3</td>
<td>8.3</td>
</tr>
<tr>
<td>JACK 0-50</td>
<td>72.0</td>
<td>72.0</td>
<td>72.0</td>
</tr>
<tr>
<td>JAE 10-40</td>
<td>78.9</td>
<td>78.9</td>
<td>78.9</td>
</tr>
<tr>
<td>BVAR 10-40</td>
<td>86.4</td>
<td>86.4</td>
<td>86.4</td>
</tr>
<tr>
<td>BIQR 10-40</td>
<td>83.5</td>
<td>83.5</td>
<td>83.5</td>
</tr>
<tr>
<td>CV 10-40</td>
<td>65.4</td>
<td>87.0</td>
<td>65.4</td>
</tr>
<tr>
<td>JACK 10-40</td>
<td>80.5</td>
<td>80.5</td>
<td>80.5</td>
</tr>
<tr>
<td>T0</td>
<td>0.0</td>
<td>32.9</td>
<td>0.0</td>
</tr>
<tr>
<td>T2.5</td>
<td>7.0</td>
<td>66.2</td>
<td>7.0</td>
</tr>
<tr>
<td>T5.0</td>
<td>24.3</td>
<td>70.8</td>
<td>24.3</td>
</tr>
<tr>
<td>T7.5</td>
<td>41.6</td>
<td>74.5</td>
<td>41.6</td>
</tr>
<tr>
<td>T10.0</td>
<td>56.1</td>
<td>77.9</td>
<td>56.1</td>
</tr>
<tr>
<td>T12.5</td>
<td>67.8</td>
<td>80.9</td>
<td>67.8</td>
</tr>
<tr>
<td>T15.0</td>
<td>77.1</td>
<td>83.7</td>
<td>77.1</td>
</tr>
<tr>
<td>T17.5</td>
<td>84.3</td>
<td>86.3</td>
<td>84.3</td>
</tr>
<tr>
<td>T20.0</td>
<td>87.5</td>
<td>87.5</td>
<td>87.5</td>
</tr>
<tr>
<td>T22.5</td>
<td>85.6</td>
<td>85.6</td>
<td>85.6</td>
</tr>
<tr>
<td>T25.0</td>
<td>83.7</td>
<td>83.7</td>
<td>83.7</td>
</tr>
<tr>
<td>T27.5</td>
<td>81.8</td>
<td>81.8</td>
<td>81.8</td>
</tr>
<tr>
<td>T30.0</td>
<td>79.8</td>
<td>79.8</td>
<td>79.8</td>
</tr>
<tr>
<td>T32.5</td>
<td>77.8</td>
<td>77.8</td>
<td>77.8</td>
</tr>
<tr>
<td>T35.0</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
</tr>
<tr>
<td>T37.5</td>
<td>73.8</td>
<td>73.8</td>
<td>73.8</td>
</tr>
<tr>
<td>T40.0</td>
<td>72.0</td>
<td>72.0</td>
<td>72.0</td>
</tr>
<tr>
<td>T42.5</td>
<td>70.1</td>
<td>70.1</td>
<td>70.1</td>
</tr>
<tr>
<td>T45.0</td>
<td>68.1</td>
<td>68.1</td>
<td>68.1</td>
</tr>
<tr>
<td>T47.5</td>
<td>65.9</td>
<td>65.9</td>
<td>65.9</td>
</tr>
</tbody>
</table>
### Table 5.7
Coverage and length of confidence intervals based on a sample of size 10

#### Normal Distribution

<table>
<thead>
<tr>
<th>Type of Interval</th>
<th>Coverage</th>
<th>Length</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5% One-sided</td>
<td>95% One-sided</td>
<td>90% Two-sided</td>
</tr>
<tr>
<td>Stand JAE 0-25</td>
<td>11.6</td>
<td>88.8</td>
<td>77.2</td>
</tr>
<tr>
<td>Boot JAE 0-25</td>
<td>8.4</td>
<td>91.2</td>
<td>82.8</td>
</tr>
<tr>
<td>Boot BVAR 0-25</td>
<td>8.0</td>
<td>91.0</td>
<td>83.0</td>
</tr>
<tr>
<td>Boot-t BVAR 0-25</td>
<td>5.2</td>
<td>94.4</td>
<td>89.2</td>
</tr>
</tbody>
</table>

#### Slash Distribution

<table>
<thead>
<tr>
<th>Type of Interval</th>
<th>Coverage</th>
<th>Length</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5% One-sided</td>
<td>95% One-sided</td>
<td>90% Two-sided</td>
</tr>
<tr>
<td>Stand JAE 0-25</td>
<td>12.6</td>
<td>90.0</td>
<td>77.4</td>
</tr>
<tr>
<td>Boot JAE 0-25</td>
<td>4.2</td>
<td>96.8</td>
<td>92.6</td>
</tr>
<tr>
<td>Boot BVAR 0-25</td>
<td>5.4</td>
<td>96.6</td>
<td>91.2</td>
</tr>
<tr>
<td>Boot-t BVAR 0-25</td>
<td>7.0</td>
<td>93.4</td>
<td>86.4</td>
</tr>
</tbody>
</table>

#### Double Exponential Distribution

<table>
<thead>
<tr>
<th>Type of Interval</th>
<th>Coverage</th>
<th>Length</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5% One-sided</td>
<td>95% One-sided</td>
<td>90% Two-sided</td>
</tr>
<tr>
<td>Stand JAE 0-25</td>
<td>12.2</td>
<td>90.8</td>
<td>78.6</td>
</tr>
<tr>
<td>Boot JAE 0-25</td>
<td>6.6</td>
<td>93.8</td>
<td>87.2</td>
</tr>
<tr>
<td>Boot BVAR 0-25</td>
<td>9.2</td>
<td>92.6</td>
<td>83.4</td>
</tr>
<tr>
<td>Boot-t BVAR 0-25</td>
<td>8.4</td>
<td>93.6</td>
<td>85.2</td>
</tr>
</tbody>
</table>
Chapter 6

Possible Extensions

In this chapter, we will briefly mention three problems where the bootstrap might be used to adaptively select a statistical procedure. The first one concerns the choice of the scaling parameter in M-estimation of a location parameter. The second one consists of generalizing the bootstrap adaptive trimmed means to the regression problem. Finally, we will conclude with the selection of the bandwidth parameter in kernel density estimation.

6.1. Bootstrap Adaptive M-estimation

Huber (1964) introduced a class of robust estimators which can be defined as the solution to the equation

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \theta}{c}\right) = 0,$$

where $c$ is a scaling parameter. Note that the contribution of each observation in the previous equation is a function of its distance from $\theta$. For this distance to be meaningful, it must first be scaled. This is different from L-estimators, such as the trimmed means, where the weight of each observation is a function of its order among the sample.

Usually, $c = ds(X)$ where $d$ is a constant and $s(X)$ is a scale equivariant statistic based on the sample $X$, often equal to the median absolute deviation from the median (MAD). This certainly is a good way to scale, but it still leaves one parameter to be determined, namely $d$. Unfortunately, the best choice of $d$ depends on the underlying distribution of the observations.

An adaptive solution consists of using the data to select the parameter $d$, or equivalently to select $c$ since $s(X)$ already depends on the data. Bell in his 1980 dissertation did this by basically choosing the value of $c$ which minimizes an estimate of the asymptotic
variance of \( \hat{\theta}(c) \), the M-estimate based on scaling parameter \( c \). He showed that under regularity conditions, the asymptotic variance of his adaptive estimator is the best possible among M-estimates with that \( \psi \) function and the underlying distribution.

The function \( \psi \) that he used is very similar to the biweight while being smoother at \( z = 1 \). It is given by \( \psi_p(x) = x \left( 1 + \frac{x^2}{2p-1} \right)^{-p} \) for \(-1 \leq z \leq 1\) and \( p > 1/2 \). In simulations with \( p = 3 \), the relative efficiency of this adaptive M-estimator to the "best known" estimator for each the sampling situations was in the neighborhood of 93-94 percent at the normal, one-wild, and slash which is better than any other estimator from a triefficiency point of view.

Yet, it might be possible to do even better by minimizing a bootstrap estimate of the finite sample variance rather than the asymptotic variance. Moreover, by using a better class of estimators than the trimmed means, such as the M-estimators based on \( \psi_{3.0} \), it might be possible to obtain a bootstrap adaptive M-estimator which would compare favorably with the best overall estimators for each of a variety of distributions, and not only among the estimators in the same class.

6.2. Bootstrap Adaptive Trimmed Mean Regression

So far, only location problems have been considered. It is an important problem for which a host of good solutions exist. Any new proposal often becomes a drop in the bucket—unless the new solution can be generalized to solve the more complex problem of regression. The class of M-estimators possesses such an advantage. On the other hand, the class of L-estimators, of which the trimmed mean is a member, could not easily be generalized until recently. The major difficulty lies in the fact that the weight given to each observation depends on the ordering of the observations. This concept is much harder to generalize in higher dimensions than that of distance on which the M-estimators depend.

Trimmed mean regression estimators have been proposed by Bickel (1973), Koenker and Bassett (1978), and Ruppert and Carroll (1980). Each of them has its drawbacks, either computational or theoretical. What seems to be a more promising approach is due
to Welsh (1987).

Basically, he uses the fact that the trimmed mean is asymptotically equivalent to
\( n^{1/2} \sum_{i=1}^{n} h(X_i) \) where \( h(x) \) is given by (4.11). Through algebra, one can rewrite \( h(x) \) as follows,

\[
h(x) = (1 - 2\alpha)^{-1} \left[ \xi_\alpha \{I(x < \xi_\alpha) - \alpha \} + x I(\xi_\alpha \leq x \leq \xi_{1-\alpha}) + \xi_{1-\alpha} \{I(x > \xi_{1-\alpha}) - \alpha \} \right] - T(\alpha, F),
\]

where \( \xi_\alpha = F^{-1}(\alpha) \). The idea is then to generalize this formula to the regression case, by
generalizing \( \xi_\alpha \) and \( \xi_{1-\alpha} \).

Following Welsh, consider \( Y_1, \ldots, Y_n \), where

\[ Y_j = x'_j \theta_0 + e_j, \quad 1 \leq j \leq n, \]

with \( \{x'_j = (x_{j1}, \ldots, x_{jp})\} \) a sequence of known \( p \) vectors, \( \theta_0 \in \mathbb{R}^p \) an unknown parameter to be estimated and \( \{e_j\} \) a sequence of i.i.d. random variables with common distribution function \( F \). For any \( \theta \in \mathbb{R}^p \), the residuals from \( \theta \) are

\[ e_j(\theta) = Y_j - x'_j \theta = e_j - x'_j(\theta - \theta_0), \quad 1 \leq j \leq n. \]

Now let \( \theta_n \) be a preliminary regression parameter estimator and let \( e_{n,1}(\theta_n) \leq e_{n,2}(\theta_n) \leq \ldots \leq e_{n,n}(\theta_n) \) denote the ordered residuals from \( \theta_n \) and for \( 0 < q < 1 \) put

\[
\xi_{nq}(\theta_n) = \begin{cases} 
  e_{n,nq}(\theta_n), & \text{if } nq \text{ is an integer,} \\
  e_{n,\lfloor nq \rfloor + 1}(\theta_n), & \text{otherwise.}
\end{cases}
\]

Then letting

\[
J_j = I\{e_j(\theta_n) \leq \xi_{n\alpha}(\theta_n)\},
\]

\[
K_j = I\{\xi_{n\alpha}(\theta_n) < e_j(\theta_n) \leq \xi_{n(1-\alpha)}(\theta_n)\},
\]

\[
L_j = I\{e_j(\theta_n) > \xi_{n(1-\alpha)}(\theta_n)\}
\]

define

\[
\tau_n = A_n^{-1} \sum_{j=1}^{n} x_j \left[ \xi_{n\alpha}(\theta_n)\{J_j - \alpha\} + Y_j K_j + \xi_{n(1-\alpha)}(\theta_n)\{L_j - \alpha\} \right],
\]

(6.2)
where $A_n^{-1}$ is any generalized inverse of $A_n = \sum_{j=1}^{n} x_j x_j' K_j$. Note that the trimmed mean regression estimator $\tau_n$ of (6.2) is indeed made up of an appropriate modification of equation (6.1).

Welsh goes on to show that an appropriate modification of Jaeckel's adaptive trimmed mean for the location problem does indeed work. This fact coupled with the way $\tau_n$ is defined leads me to conjecture that the theory developed in chapters 3 and 4 should also work in this problem.

It is interesting to note that bootstrap regression trimmed means have been studied in a Monte Carlo simulation by de Jongh and de Wet (1985). They used the regression trimmed mean of Koenker and Bassett (1978). In that study, an adaptive regression trimmed mean similar to Jaeckel's estimator did better, overall, than a bootstrap adaptive trimmed mean based on minimizing an estimate of mean squared error, although the latter did much better for the normal distribution. It would certainly be very interesting to see how a bootstrap adaptive regression trimmed mean would do using Welsh's $\tau_n$ and the information that we gathered for the location problem in chapter 5.


In this section we present a third problem involving a smoothing parameter where an adaptive choice is very useful: kernel density estimation. The kernel density estimator goes back to Rosenblatt (1956) and Parzen (1962). It is given by

$$f_{n,c}(x) = \frac{1}{cn} \sum_{i=1}^{n} k\left(\frac{x - X_i}{c}\right),$$

where $k$ is a density and $c$ is a smoothing parameter. Density estimation is the subject of numerous books such as Tapia and Thompson (1978), Devroye and Györfi (1985), and Silverman (1986).

To obtain consistency, the smoothing parameter must converge to 0 at a certain rate, with the "best" sequence depending on the underlying density $f$. Best is in quotes because no optimizing criterion was specified. But whatever criterion is used, the best result still
depends on the true density. For small sample sizes, the best smoothing parameter is not known. Small values of the smoothing parameter leads to an estimate with low bias and high variance, whereas large values give rise to low variance and high bias. It therefore makes sense to, once again, use the data to choose the smoothing parameter \( c \).

The usual method is based on cross-validation. See Rudemo (1982) and Bowmann (1984). The idea consists of choosing \( c \) so as to minimize

\[
L(c) = \prod_{i=1}^{n} f_{n,c}^i(X_i),
\]

where

\[
f_{n,c}^i(x) = \frac{1}{(n-1)c} \sum_{j=1 \atop j \neq i}^{n} k \left( \frac{x - X_i}{c} \right).
\]

Instead, one might choose \( c \) which minimizes a bootstrap estimate of the (finite sample) mean square error of \( f_{n,c} \) either at a fixed point \( x \) or the average of the mean squared error at each observation.

Of course, when dealing with the bootstrap in the context of density estimation, one must exercise care in the bootstrapping step because otherwise it might not even be consistent. See Romano (1988).
Chapter 7

Conclusion

Numerous estimators in different problems, ranging from the estimation of a finite dimensional location parameter to the infinite dimensional estimation of a density, are only specified up to a constant which will be called the smoothing parameter. In most cases, the optimal choice of that parameter is a function of the underlying distribution of the observations and of the sample size. Using a fixed value will lead to good or bad estimation depending on the particulars of the sample at hand. Consequently, the smoothing parameter is often chosen adaptively, that is, using the data.

One way to do this is by choosing the value of the smoothing parameter such that the estimator optimizes an estimate of a certain criterion. Up until recently, estimates of characteristics of the finite sample distribution of most estimators were hard to get because that distribution is unknown. On the other hand, in some cases, closed form estimates of characteristics of the asymptotic distribution are known, such as the asymptotic variance of the trimmed means. Thus, adaptive methods based on optimizing characteristics of the asymptotic distribution of the estimators were introduced. But the bootstrap can often be used to estimate characteristics of the finite sample distribution. By using them, it is possible to optimize estimates of a finite sample criterion, often a more meaningful criterion. Not only is it a viable alternative to using an asymptotic criterion, but sometimes it is the only alternative as no closed form estimator of characteristics of the asymptotic distribution exists.

This dissertation illustrates this approach by looking at trimmed means in the location problem. The theory in chapters 3 and 4 shows that if the estimators vary smoothly in the smoothing parameter and the distribution of the observations, then it is possible to
use the bootstrap to adaptively select the smoothing parameter.

The actual criterion used in adaptively selecting the smoothing parameter is very much problem dependent. In the case of trimmed means, minimizing the variance and the interquartile range of the finite sample distribution both worked well. In general, one wants to use a finite sample criterion such that the adaptive choice will converge in probability to the best smoothing parameter from an asymptotic point of view. By using a finite sample criterion, one is likely to obtain good finite sample results. By using an adaptive choice which converges to the best asymptotic value of the smoothing parameter, the adaptive statistical procedure will do asymptotically as well as possible among that class of estimators. In this example, both of these goals were achieved by the bootstrap adaptive trimmed means, as illustrated by the asymptotic theory of chapter 3, and the results of the simulations in chapter 5.

The first message that we want to convey is that the bootstrap can successfully be used to construct adaptive statistical procedures with good finite sample results and good asymptotic properties.

The second message is rather old and well-known, but can never be emphasized enough, especially in view of the availability of methods to deal with the problem. Do not make inference statements based on an adaptive procedure by assuming that the adaptively chosen parameter was a priori fixed to the adaptive choice. In the example that was investigated, this translates into not using a confidence interval that does not use the adaptive nature of the estimator in the determination of the bounds of the intervals. The confidence interval based on the asymptotic normality of the estimator does not take into account the adaptivity of adaptive trimmed means and lead to coverage probabilities consistently at least 12% lower than the claimed 90% at least for the four sampling situations studied. On the other hand, bootstrap confidence intervals that share the same asymptotic properties consistently did better in small samples. Such intervals are based on the finite sample distribution of the same adaptive statistic, but for data whose distribution is different, namely, \( \hat{F}_n \) instead of \( F \). Hence, the adaptivity of the estimators is clearly
Thirdly, this dissertation illustrates a particular approach to adaptive estimation. In this approach, a class of estimators indexed by a one-dimensional parameter is selected. The class should be large enough to contain good estimators for a large family of distribution functions, but small enough to guarantee that the estimators behave well uniformly over the class, a necessary condition to use the bootstrap. Such an approach is an alternative to the fully efficient adaptive estimators which asymptotically are the best estimators for a large family of distributions that satisfy certain regularity conditions. The class of estimators from which these estimators are adaptively chosen from is indexed by a function instead of a parameter. By, in essence, estimating only one parameter instead of a function, the adaptive trimmed means are much simpler and have been shown to do rather well over four widely different sampling situations for samples of size as small as 10. The fully efficient adaptive estimators have not been shown to do as well for very small sample sizes.

Finally, it is hoped that the bootstrap adaptive trimmed means introduced in this dissertation will not simply be viewed as yet two more estimators of location. Instead, they should be viewed as a manageable first step in establishing whether bootstrap adaptive procedures in more difficult problems, such as regression or density estimation, are indeed good and justified asymptotically. Evidently, much more work remains.
Bibliography


To appear.


