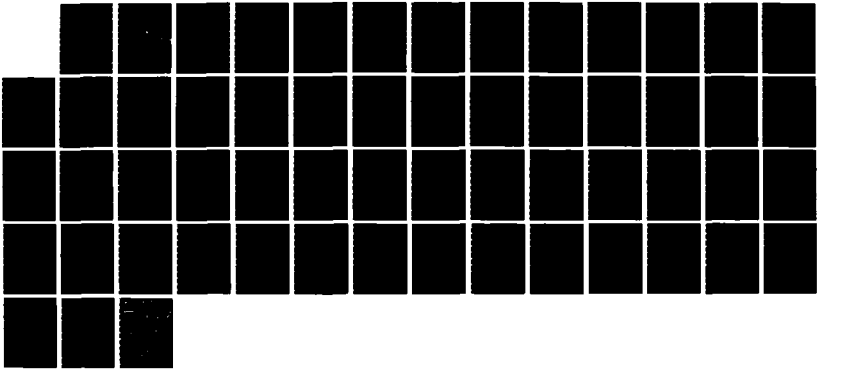
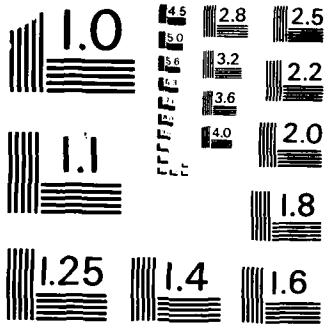
AD-A175 149 AN ANALYSIS OF THE BOOTSTRAP METHOD FOR ESTIMATING THE MEAN SQUARED ERROR OF STATISTICAL ESTIMATORS(U) NAVAL POSTGRADUATE SCHOOL MONTEREY CA W CORTES-COLON SEP 86 1/1

UNCLASSIFIED F/G 12/1 NL





NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

AN ANALYSIS OF THE BOOTSTRAP METHOD
FOR ESTIMATING THE MEAN SQUARED ERROR
OF STATISTICAL ESTIMATORS

by

William Cortes-Colon

September 1986

Thesis Co-Advisors: Donald R. Barr
T. Jayachandran

Approved for public release; distribution is unlimited.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Naval Postgraduate School 6b OFFICE SYMBOL (*If applicable*) Code 55	7a. NAME OF MONITORING ORGANIZATION Naval Postgraduate School
6c. ADDRESS (*City, State, and ZIP Code*) Monterey, California 93943-5000	7b. ADDRESS (*City, State, and ZIP Code*) Monterey, California 93943-5000
8a NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (*If applicable*)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (*City, State, and ZIP Code*)	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO / PROJECT NO / TASK NO / WORK UNIT ACCESSION NO

11 TITLE (*Include Security Classification*)
AN ANALYSIS OF THE BOOTSTRAP METHOD FOR ESTIMATING THE MEAN SQUARED ERROR OF STATISTICAL ESTIMATORS

12 PERSONAL AUTHOR(S)
Cortes-Colon, William

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (*Year, Month, Day*)	15 PAGE COUNT
Master's Thesis	FROM _____ TO _____	1986, September	58

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB-GROUP	BOOTSTRAP, NON-PARAMETRIC, BOOTSTRAP ESTIMATOR

19 ABSTRACT (*Continue on reverse if necessary and identify by block number*)

One of the most crucial problems in theoretical and applied statistics is to determine the precision of the estimates produced by different statistical estimators. This problem is greatly increased when the population parametric characteristics are not known. Parallel to this problem is that of deciding how large (or small) the sample population must be in order to obtain a desired precision within certain range.

There are several non-parametric methods to approach the first problem. The BOOTSTRAP method (Efron, 1979) is one of these approaches and the one of interest in this thesis. With this method, one could improve the precision of the estimates and gain information about the distributional characteristics of statistical estimators. The bootstrap method has been amply compared with other methods; the results show that the bootstrap method often produces more precise estimates (i.e. with smaller mean squared error) than competitors such as the JACKNIFE, SECTIONING and CROSS-VALIDATION. However, the results that have been obtained are based on large sample sizes and large numbers of "bootstrap"

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION	
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED	
22a NAME OF RESPONSIBLE INDIVIDUAL Donald R. Barr	22b TELEPHONE (*Include Area Code*) (408) 646-2663	22c OFFICE SYMBOL Code 55 Bn

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

19. ABSTRACT

replications.

This thesis analyzes the behavior of the BOOTSTRAP method when the number of bootstrap replications is small. It tries to identify any tradeoffs between sample size and the number of bootstrap replications required to attain a desired precision in the estimates produced in several particular situations. One of the goals is to produce graphical displays that will indicate to the experimental statistician the price that must be paid in the precision of the estimates, obtained with the bootstrap method, when sample size is small, and the number of bootstrap replications to use in this situation.

Approved for public release; distribution is unlimited.

An Analysis of the Bootstrap Method for Estimating the Mean Squared Error of Statistical Estimators

by

William Cortes Colon
Captain (P), United States Army
B.S., University of Andorra, 1972
M.S., University of Navarra,Spain, 1975

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1986

Author: _____
William Cortes Colon

Approved by: _____
Donald R. Barr, Thesis Advisor

_____
Toke Jayachandran, Co-Advisor

_____
A. R. Washburn, Chairman,
Department of Operations Research

_____
Kneale T. Marshall,
Dean of Information and Policy Sciences

ABSTRACT

One of the most crucial problems in theoretical and applied statistics is to determine the precision of the estimates produced by different statistical estimators. This problem is greatly increased when the population parametric characteristics are not known. Parallel to this problem is that of deciding how large (or small) the sample population must be in order to obtain a desired precision within certain range.

There are several non-parametric methods to approach the first problem. The BOOTSTRAP Method (Efron, 1979) is one of these approaches and the one of interest in this thesis. With this method, one could improve the precision of the estimates and gain information about the distributional characteristics of statistical estimators. The bootstrap method has been amply compared with other methods; the results show that the bootstrap method often produces more precise estimates (i.e. with smaller mean squared error) than competitors such as the JACKNIFE, SECTIONING and CROSS-VALIDATION. However, the results that have been obtained are based on large sample sizes and large numbers of "bootstrap" replications.

This thesis analyzes the behavior of the BOOTSTRAP method when the number of bootstrap replications is small. It tries to identify any tradeoffs between sample size and the number of bootstrap replications required to attain a desired precision in the estimates produced in several particular situations. One of the goals is to produce graphical displays that will indicate to the experimental statistician the price that must be paid in the precision of the estimates, obtained with the bootstrap method, when sample size is small, and the number of bootstrap replications to use in this situation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

A. BACKGROUND

One of the most common problem in applied statistics is the estimation of an unknown parameter θ. Once the statistician has decided on the model having one or more parameters to be estimated and has selected *the estimator* (i.e., m.l.e., least-square estimator, etc.) that will be used to obtain the estimates, the second problem that he or she faces is how to estimate the accuracy of these estimates. There are several ways of measuring the accuracy or the error of statistical estimators. In this thesis, the measure of statistical error will be defined to be the mean squared error (MSE) of the estimators; i.e. the variance plus the bias-squared of θ^h (where θ^h represents the estimator of the parameter θ. In Appendix A the reader will find a list of special notations used in this thesis) :

$$MSE(\theta^h) = E[(\theta^h - \theta)^2] = Var(\theta^h) + [BIAS(\theta^h)]^2 \qquad (1.1)$$

When the practitioner is dealing with samples obtained from populations for which the distributional characteristics are known, classical statistical theory provides an answer to the second problem that the statistician faces. This is true since, at least in theory, the variance and the bias of most statistical estimators can be calculated analytically. However, the difficulty of analytically deriving the MSE of some statistical estimator increases as the mathematical definition of the estimator becomes more complicated. When this is the case or when the practitioner does not actually know the probability distribution, say F, from which the sample was obtained, then the MSE of the estimators must be estimated.

There are several non-parametric methods for estimating the bias and the variance of an estimator of interest. The most common ones are the Quenoille-Tukey JACKNIFE method, CROSS-VALIDATION, and SECTIONING; the Jacknife being the most commonly used of the three approaches. Efron and Gong [Ref. 1] and Miller [Ref. 2] provide an excellent exposition of the first two methods and Lewis gives a good introduction and analysis of the later (See [Ref. 3]).

In recent years, Efron [Refs. 1,4], has developed another, rather intriguing non-parametric methodology for estimating the MSE of any statistic. This method, called the BOOTSTRAP, is simple and has been shown by Efron to be a powerful statistical tool that can be applied even in complex situations (See Efron, [Ref. 5] and [Ref. 6]). This method, as shown in this thesis, is a good approach for estimating the precision of a statistical estimator used in a given model. It also gives information about the distributional characteristics of the estimator used. Efron and Gong [Ref. 1] and Tibshirani [Ref. 7] have conducted intensive analyses of this new method and have compared it with the other non-parametric methods mentioned above. Surprisingly for some authors, the BOOTSTRAP has been shown to produce estimates with much more precision (sometimes up to twenty percent lower variance, for example) than the JACKNIFE and CROSS-VALIDATION estimators. As an example, Efron [Ref. 4: Section 3], has shown that the BOOTSTRAP methodology correctly estimates, asymptotically, the variance of the sample median, a case where the JACKNIFE is known to fail. As in the case of the sample median, it is known that the JACKNIFE collapses for non-smooth statistics; however, the BOOTSTRAP seems to produce accurate estimates even in these cases.

B. THE GENERAL PROBLEM

Suppose that the realization $x_1, x_2, \ldots, x_n$ of a random sample $X_1, X_2, \ldots, X_n$ has been observed, and that $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.), having a probability distribution F. In practice, the distribution F is probably unknown and the problem is to estimate the value of some parameter of interest, such as the mean, variance, or median. This is done using a sample of size n with some estimator of $\theta(F)$, say $\theta^h(F)$. The basic idea of the BOOTSTRAP method is very simple, at least in principle:[1] having observed $x_1, x_2, \ldots, x_n$, *construct the sample empirical probability distribution*, F^h, by putting mass 1/n at each observation $x_1, x_2, \ldots, x_n$. Now, fixing F^h, draw a random sample of size n *with replacement* from F^h. This sample will be called a *bootstrap random sample* and will be denoted by

$$X^* = (X^*_1, X^*_2, \ldots, X^*_n) \tag{1.2}$$

[1]The BOOTSTRAP methodology will be analyzed in more detail in Chapter 2.

and then $X^*_i \sim_{iid} F^h$. Then the task is to estimate the distribution of $\theta(F)$ by the distribution of $\theta^*(F^h)$, where $\theta^*(F^h)$ denotes the value of the parameter of interest based on the bootstrap mechanism. This mechanism proceeds as follows : keeping F^h fixed, draw a bootstrap sample and calculate $\theta^*(F^h)$; do this a large number B of times obtaining $\theta^*_1(F^h)$, $\theta^*_2(F^h)$, . . . , $\theta^*_B(F^h)$. The resultant (sample) distribution of θ^* is called the *bootstrap distribution* F^{h*} . Once F^{h*} is obtained, then any specific feature of this distribution, such as expected value of θ^* , $E_*(\theta^*)$ or the variance of θ^* , $Var_*(\theta^*)$, could be obtained. (In this thesis, notation like "E_* ", "Var_* ", "S^{*2} ", "X^* " , etc., indicates calculations relating to the *conditional bootstrap distribution* of X^* , with the vector of random variates *X* and hence F^h , fixed.[2]). Theoretically, then, the bootstrap idea could be used to estimate the expected value, the variance, and the mean squared error of any estimator, given a sample that comes from an unknown probability distribution F.

As mentioned earlier, Efron (See [Ref. 4]) has shown that this method is often more precise than other non-parametric methods for assessing statistical accuracy. However, the experimentation done in the past using this method relied on *a large* number B of bootstrap replications; i.e, a large sample on θ^*. In some cases, it can be shown (see Chapter 2, for the case of $Var_*(\theta^*)$) that as $B \rightarrow \infty$, the variance of θ^* based on F^h is equal to the variance of the estimator θ based on F . But, how large must B be in order to obtain estimates that are accurate or to obtain estimators with a small MSE is a question to be answered. Also, what is the tradeoff between the sample size n and the number B of bootstrap replications ?

The purpose of this thesis is then twofold : first, to analyze the bootstrap performance as the number B of replications increases, starting from a small B. The second, also of great interest, is to study the relationship between the sample size n and the number B in the estimation of the MSE of the estimator using the bootstrap mechanism.

C. ORGANIZATION

There are several methods of dertermining the bootstrap distribution of an estimator $\theta^*(F^h)$, two of which will be analyzed in this thesis.[3] The first is by direct

---

[2]As it will be shown in the next chapter, this is a critical feature of the BOOTSTRAP method: the vector of random variates *X* and F^h must be fixed through the process.

[3]A third method involves making Taylor series expansion to obtain the

theoretical calculations (this is usually the most difficult approach). The second relies on Monte Carlo approximations to the bootstrap distribution: repeated realizations of X^* are generated by taking random samples of size n from F^h , say x^{*1} , x^{*2} , . . . , x^{*B} and the histogram of the corresponding values $\theta^*_1(F^h)$, $\theta^*_2(F^h)$, . . . , $\theta^*_B(F^h)$ is constructed as an approximation to the actual bootstrap distribution (See [Ref. 1: Section 2]). These two methods are of interest in the second chapter. In the last section of Chapter Two, the different statistical experiments conducted for this thesis are explained in detail. In Chapter Three, the results from these experiments are presented and analyzed, and the problem of using the bootstrap approach in linear regression problems is also discussed. Conclusions are presented in the last chapter. There, one of the points of interest is to discuss the main disadvantage of the bootstrap methodology : the computer time required to implement this method when Monte Carlo simulation is used. In Appendix B, the FORTRAN software that was designed to run the experiments discussed in this thesis will be explained and the code is listed. This computer program is user friendly and can be used to estimate the bootstrap distribution of eight different estimators. Finally in Appendix C, the reader can see some tables that give a good idea about how large (or small) B and n can be in order to obtain a desired precision on the estimates of parameters of given populations F.

---

approximate mean and variance of the bootstrap distribution F^* . See Ref.4, Section 5.

II. THE BOOTSTRAP METHOD

A. A DESCRIPTION OF THE METHOD

As mentioned earlier, the Bootstrap methodology is, in principle, simple. Also, recall that in this thesis the problem of interest is to study how this method performs in estimating the MSE of some statistical estimators, and how the MSE behaves as the number B of bootstrap replications and the sample size n change.

Suppose that the data of interest consist of a random sample $X = (X_1, X_2, \ldots, X_n)$ of size n, from an unspecified probability distribution F on the real line. The X_i may be real valued, two dimensional, or take values in a more complicated space, but this will not affect the theory, see Efron [Ref. 2]. Thus, it is assumed that

$$X_1, X_2, \ldots, X_n \sim_{iid} F. \tag{2.1}$$

The problem is now to estimate the probability distribution of a specific estimator of a parameter $\theta(F)$, say $\theta^h(F)$. The probability distribution of $\theta^h(F)$ could be approximated by the following algorithm (See Efron [Ref. 1: Section 2]) :

(1) given that the realization of X has been observed, say $X_i = x_i$, i = 1, 2,..., n,

(2) construct the sample probability distribution F^h , by putting mass 1/n at each point $x_1, x_2, \ldots, x_n$,

(3) keeping x_i and F^h fixed, draw *with replacement* a random sample of size n from F^h , and call this the bootstrap sample; i.e., $X^*_i = x^*_i$, where $X^*_i \sim_{iid} F^h$, so

$$P(X^*_i = x_j \mid X = x) = 1/n \ , \tag{2.2}$$

(4) the distribution of $\theta^h(F)$ can be approximated by a sample on $\theta^*(F^h)$; then, a measure of accuracy could be assigned to $\theta^*(F)$ base on $\theta^*(F^h)$.

As mentioned earlier, the distribution of some estimators $\theta^*(F^h)$ might be calculated analytically.

1. *Direct Analytical Calculations*

An attempt is now made to calculate some parameters of interest of the distribution of X^*_i. Assuming the conditions shown in expressions (2.1) and (2.2), the expected value of X^*_i, given X, could be calculated as follows :

$$E_*(X^*_i) = E(X^*_i \mid X = x) = \sum_j x_j \; P(X^*_i = x_j \mid X = x) , \qquad (2.3)$$

where j = 1, 2,..., n. From (2.2), this is equal to :

$$E_*(X^*_j) = \sum_j (x_j / n) = \bar{X} \qquad j = 1, 2,..., n , \qquad (2.4)$$

which is the sample mean of the original sample X. Then from (2.4), the unconditional expected value of X^*_i is :

$$E(X^*_j) = E[E_*(X^*_j \mid X)] = E(\bar{X}) = \mu_X \qquad j = 1, 2,..., n . \qquad (2.5)$$

Thus, the unconditional expectation of X^*_j is equal to the mean of the population from which the original sample was obtained. (Note, from this point on all summation signs go from 1 to n, unless otherwise specified, and E_* , Var_* , etc., are conditional, give X .)

Likewise, the unconditional variance of X^* could be derived from the conditional variance of X^* :

$$Var_*(X^*_i) = E_*[(X^*_i - E(X^*_i \mid X = x))^2] . \qquad (2.6)$$

Using (2.5) this expression is equivalent to :

$$\begin{aligned} Var_*(X^*_i) &= E[(X^*_i - \bar{X})^2 \mid X] \qquad (2.7) \\ &= E_*(X^{*2}_i) - \bar{X}^2 \\ &= \sum_i (X^2_i / n) - \bar{X}^2 \\ &= \sum_i (X_i - \bar{X})^2 / n \end{aligned}$$

By definition of the sample variance, S^2_X , then

$$Var_*(X^*_i) = (n-1)/n \; S^2_X \qquad (2.8)$$

Now, unconditionally

$$\begin{aligned} \text{Var}(X^*_i) &= E(\text{Var}_*(X^*)) + \text{Var}[E_*(X^*)] \\ &= E[\textstyle\sum_i (X^2_i / n) - \bar{X}^2] + \text{Var}(\bar{X}) \\ &= E[(n-1)/n\, S^2_X] + \sigma^2_X / n \\ &= (n-1)/n\, E(S^2_X) + \sigma^2_X / n \\ &= (n-1)/n\, \sigma^2_X + \sigma^2_X/n \\ &= \sigma_X^2 \end{aligned} \tag{2.9}$$

Therefore, the variance (unconditional) of X^*_i is the same as the variance of X_i. The covariance between X^*_i and X^*_j has a very important impact on the bootstrap methodology, primarily when the bootstrap distribution of $\theta^*_i(F^h)$ is approximated by Monte Carlo simulation (see next section).

Conditionally (given X), the covariance between X^*_i and X^*_j is as follows :

$$\text{Cov}_*(X^*_i, X^*_j) = E_*[(X^*_i - E_*(X^*_i))\,(X^*_j - E_*(X^*_j))] \ . \tag{2.10}$$

From (2.5), this is

$$\begin{aligned} \text{Cov}_*(X^*_i, X^*_j) &= E_*[(X^*_i - \bar{X})\,(X^*_j - \bar{X})] \\ &= E_*(X^*_i\, X^*_j) - \bar{X}^2 \end{aligned} \tag{2.11}$$

Now conditionally, given $X = x$, the joint distribution of (X^*_i, X^*_j) is *uniform* over the points $(x_1, x_2, \ldots, x_n) \times (x_1, x_2, \ldots, x_n)$ and this implies that $(X^*_i\, X^*_j) = (x_k x_l)$ with probability $1/n^2$. Then

$$\begin{aligned} E_*(X^*_i\, X^*_j) &= \textstyle\sum_i \sum_j (x_i\, x_j) / n^2 \qquad\qquad i \neq j \\ &= (1/n^2)\,(\textstyle\sum_i x_i)^2 = \bar{x}^2. \end{aligned} \tag{2.12}$$

Finally, the conditional covariance between X^*_i and X^*_j is

$$\text{Cov}_*(X^*_i, X^*_j) = \bar{X}^2 - \bar{X}^2 = 0 \ . \tag{2.13}$$

Now, to derive the unconditional covariance between X^*_i and X^*_j, it will be convenient to use the result obtained in equation (2.13). To use (2.13), it must be shown that the following equality holds:

$$\mathrm{Cov}(X^*_i,X^*_j) = E[\mathrm{Cov}_*(X^*_i,X^*_j)] + \mathrm{Cov}[E_*(X^*_i), E_*(X^*_j)]. \tag{2.14}$$

To show this, notice that the conditional covariance can be defined as

$$\begin{aligned}\mathrm{Cov}(X,Y|Z) &= E_{(x,y|z)}[(XY - E(X|Z)E(Y|Z))|Z] \\ &= E_{(x,y|z)}(XY|Z) - [E(X|Z)E(Y|Z)] .\end{aligned} \tag{2.15}$$

Then

$$\begin{aligned}E_Z[\mathrm{Cov}(X,Y|Z)] &= E_Z[E_{(x,y|z)}(XY|Z) - \{E(X|Z)E(Y|Z)\}] \\ &= E_Z[E_{(x,y|z)}(XY|Z)] - \{E_Z[E(X|Z)]E_Z[E(Y|Z)]\} - \\ &\quad - E_Z[E(X|Z)E(Y|Z)] + \{E_Z[E(X|Z)]E_Z[E(Y|Z)]\} \\ &= \mathrm{Cov}(X,Y) - \mathrm{Cov}[E(X|Z),E(Y|Z)].\end{aligned} \tag{2.16}$$

Therefore,

$$\mathrm{Cov}(X,Y) = E_Z[\mathrm{Cov}(X,Y|Z)] + \mathrm{Cov}[E(X|Z),E(Y|Z)]. \tag{2.17}$$

With this in mind, the unconditional covariance could finally be computed by using (2.15). Now, the portion inside the brackets of the first term of the right hand side of equation (2.14) was shown in (2.13) to be equal to zero. Then, using expression (2.5), equation (2.14) reduces to–

$$\mathrm{Cov}(X^*_i,X^*_j) = \mathrm{Cov}(\bar{X},\bar{X}) = \mathrm{Var}(\bar{X}) = \sigma^2_X/n \ , \tag{2.18}$$

and from (2.18), the correlation coefficient is given by

$$\rho(X^*_i,X^*_j) = 1/n = P[X^*_i = X_j] \tag{2.19}$$

Comparing equations (2.13) and (2.18) it could then be stated that the bootstrap samples are (conditionally) independent as long as X is held fixed.

It is possible now to derive the distributional characteristics of some statistical estimators based on the distribution of X^*_i. In doing this, it is assumed that the original sample X is fixed and these derivations are conditional. For example, the expected value and the variance of $\bar{X}^*$ (the bootstraped sample mean) are obtained as follows: using equation (2.5)

$$E_*(\bar{X}^*) = \bar{X} \ , \tag{2.20}$$

so unconditionally, the expected value of the bootstrap sample mean is

$$E(\bar{X}^*) = E(\bar{X}) = \mu_X \ . \tag{2.21}$$

The conditional variance of the bootstrap sample mean is

$$\begin{aligned} Var_*(\bar{X}^*) &= (1/n^2)Var_* [\ \textstyle\sum_i (X^*_i)] \\ &= (1/n^2) [\textstyle\sum_i Var_*(\bar{X}^*_i) + (n(n-1)/2)Cov_*(X^*_i, X^*_j)] \ . \end{aligned} \tag{2.22}$$

From equation (2.13), the conditional variance is then

$$\begin{aligned} Var_*(\bar{X}^*) &= (1/n^2)[\textstyle\sum_i Var_*(X^*_i)\ ] \\ &= (1/n^2)\ [n\ Var_*(X^*_i)] \ . \end{aligned} \tag{2.23}$$

Using equation (2.8), finally

$$Var_*(\bar{X}^*) = (n-1)/n^2\ S^2_X \ . \tag{2.24}$$

With this expression, the unconditional variance of $\bar{X}^*$ is given by

$$Var(\bar{X}^*) = E[Var_*(\bar{X}^*)] + Var[E_*(\bar{X}^*)] \ . \tag{2.25}$$

From equation (2.5), and (2.20)

$$\begin{aligned} Var(\bar{X}^*) &= E[(n-1)/n^2\ S^2_X] + Var(\bar{X}) \\ &= (n-1)/n^2\ \sigma^2_X + \sigma^2_X/n \\ &= (2n-1)/n\ Var(\bar{X}) \end{aligned}$$

As mentioned earlier, equation (2.24) is the one of interest when one wants to apply the bootstrap mechanism to obtain the variance of $\bar{X}^*$. Notice that as $n \rightarrow \infty$,

$$Var_*(\bar{X}^*) \rightarrow Var(\bar{X}) \tag{2.26}$$

strongly (strong law of large numbers), but this is not the case for the unconditional variance of $\bar{X}^*$, where as $n \rightarrow \infty$,

$$\text{Var}(\bar{X}^*) \rightarrow 2\text{Var}(\bar{X}) . \tag{2.27}$$

It is now possible to define an estimator for the MSE of the mean of a population based on X^*:

$$\begin{aligned} \text{MSE}_*(\bar{X}^*) &= \text{Var}_*(\bar{X}^*) + [E_*(\bar{X}^* - E_*(\bar{X}^*)]^2 \\ &= \text{Var}_*(\bar{X}^*) + [\text{Bias}_*(\bar{X}^*)]^2 \end{aligned} \tag{2.28}$$

In the same manner, the MSE of any estimator could be derived. However, it is easy to see that as the mathematical definition of the estimator gets more complicated, this procedure can become very tedious. This is why it is desired to estimate the bootstrap distribution of the estimator by simulation rather than analytically.

2. *Monte Carlo Simulation*

The algorithm presented in Chapter II, Section A, could be expanded to allow Monte Carlo simulation to approximate the bootstrap distribution of $\theta^*(F^h)$. As before (See Efron [Ref. 2: Section 2]):

(1) given that the realization of the random vector X has been observed, say $X_i = x_i$ for i= 1, 2,..., n ;

(2) construct the sample probability distribution F^h , by giving a mass 1/n at each point $x_1 , x_2 , \ldots , x_n$,

(3) keeping x_i (and thus, F^h) fixed, draw *with replacement* a random sample of size n from F^h , and call this a bootstrap sample;

(4) from this random sample, compute the bootstrap replication, $\theta_i^*(F^h)$; i.e, compute the value of the desire statistic based on the sample from F^h . Then,

(5) do steps (3) and (4) a "large" number B of times. In this way one obtains independent bootstrap replications of $\theta^*(F^h)$, say $\theta^*_1(F^h)$, $\theta^*_2(F^h)$,..., $\theta^*_B(F^h)$;

(6) now, approximate the variance of $\theta^*(F^h)$ by the sample variance

$$\text{Var}_*{}^h[\theta^*(F^h)] = \sum_i [\theta^*_i(F^h) - \bar{\theta}^*(F^h)]^2 / (B - 1) , \tag{2.29}$$

where i= 1, 2,..., B, and

$$\bar{\theta}^*(F^h) = \sum_i \theta^*_i(F^h) / B . \tag{2.30}$$

The MSE of $\theta^*(F^h)$ may be estimated by

$$MSE_*^h(\theta^*(F^h)) = Var_*^h[\theta^*(F^h)] + [BIAS_*^h(\theta^*(F^h)]^2 . \qquad (2.31)$$

It will be seen in Chapter Three that as B and n get large $MSE_*^h(\theta^*(F^h))$ approaches zero. A problem in using the bootstrap is the choice of B, and we consider this in Chapter Three.

This bootstrap simulation procedure was carried out to study the effect of possible choices of B, in terms of the estimated MSE of several estimators. The reader will see, in the next chapter, that the choice of B should depend on the sample size n, the specific estimator under consideration and the structure of the population from which the sample was obtained.

a. The Statistical Experiment

In this thesis, various experiments were conducted to study the problem of selecting B. The main idea behind these experiments was to select some well known probability distributions and some parametric estimators for which the distributional characteristics are well known. Then the MSE of these estimators could be determined theoretically. Therefore, one could compare this true MSE with the estimated MSE of the estimators obtained using the bootstrap mechanism.

The critical part of the experiment was to design an effective computer code to perform the Monte Carlo simulation. The FORTRAN program developed to carry out the simulation reported here is listed in Appendix B. This program was used to analyze the performance of eight different estimators based on the bootstrap methodology. These were the sample mean, variance (three different estimators), coefficient of correlation, coefficient of variation, the five-percent trimmed mean, and the median.

The simulation runs as follows (See Appendix B):

(1) n random variates, for up to 8 values of n, are first generated representing a random sample from a population F. (In the simulation a total of N random variables are first generated, then sectioned into samples of sizes n_i where i= 1, 2, ..., 8.)

(2) For each subsample of size n , a *bootstrap function* is called to generate a bootstrap sample from the original sample. Then, the *estimator function* is

called to produce a desired estimate. This step is repeated until B bootstrap samples from the original sample are obtained.

(3) After the B estimates have been obtained, the *statistics function* is called to calculate the mean of these estimates, this number is one of the $\theta^*_i(F^h)$.

(4) In order to improve the precision of the simulation process, steps (2) and (3) are replicated M times. Then, the process will produce a total of $(N \times M)/ n$ estimates. From these estimates, a box-plot is constructed and estimates, including MSE, are calculated.

In the next chapter some of the results obtained from this simulation process are analyzed.

III. APPLICATION OF THE BOOTSTRAP METHOD : SOME RESULTS

A. THE MEAN, VARIANCE AND THE COEFFICIENT OF VARIATION OF EXPONENTIAL RANDOM VARIATES

The first experiment conducted was intended to analyze the bootstrap mechanism in estimating the MSE of the estimators for the mean, variance and coefficient of variation of a sample coming from a population of exponential random variates with parameter $\lambda = 1$. The population coefficient of variation is defined as:

$$CV(X) = \sigma_X/\mu_X \tag{3.1}$$

In the Exponential(1) case, the mean, variance and the coefficient of variation have the same value of 1. With this first fact in mind, the MSE of sample mean, as an example, is defined using (2.21) and (2.28) as:

$$MSE(\bar{X}^*) = Var(\bar{X}^*) + [E(X^* - \mu_X)]^2 \ . \tag{3.2}$$

Conditionally, from (2.26), an estimate of (3.2) is:

$$MSE_*^h(\bar{X}^*) = [(n-1)/n^2 \, S_X^2] + [E_*(X^* - 1)]^2 \ . \tag{3.3}$$

In the same manner, the MSE for the variance and coefficient of variation could be estimated. These estimates were obtained using the algorithm described in the preceding section. The sample sizes for this experiment were: n = 10, 20, 25, 40, 50, 70 ,100, 140. Each estimator was *bootstraped* using B = 5, 8, 10, 15, 20, 25, 40, 60, 100, 140, and 500. Figures 3.1, 3.2 and 3.3 below, show how the MSE_*^h for the mean, variance and coefficient of variation respectively decreases as both n and B increases.

A remarkable feature of these plots is that the MSE_*^h of the bootstrap sample variance (Figure 3.2) decreases much faster as the sample size increases than when B increases. Observe the big jump in the MSE_*^h when n goes from 10 to 40 relative to that of B going from 5 to, say, 40: the jump is much greater in the former.

Another observation of interest is that the MSE_*^h of the estimates decreases as B increases, but beyond a certain threshold very slowly. Indeed, the decrease in MSE_*^h

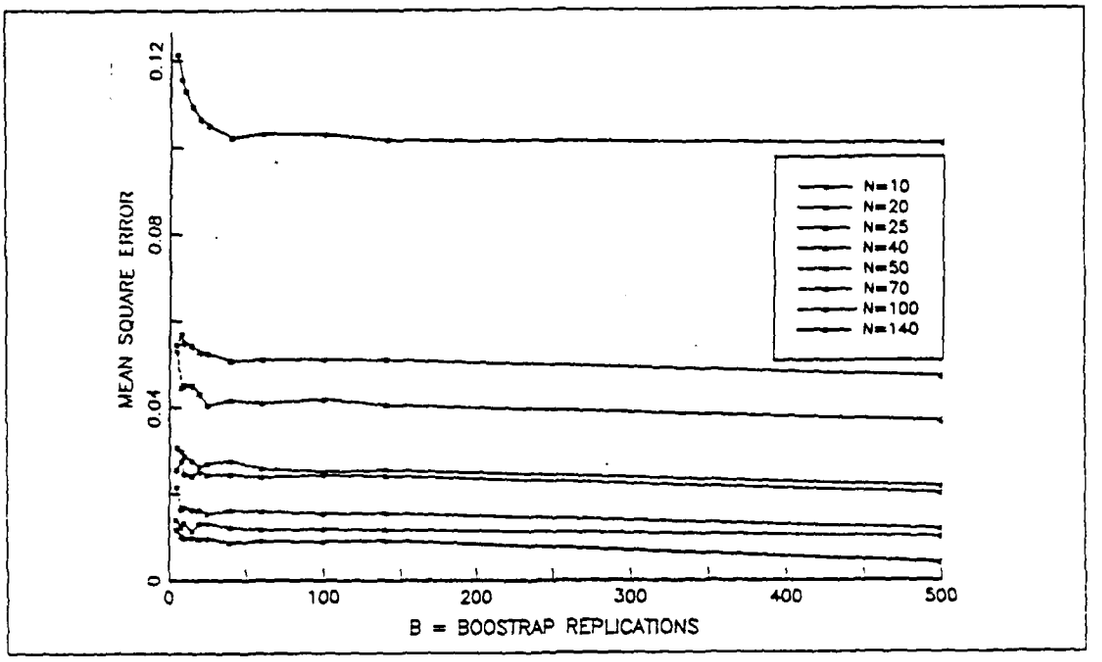


Figure 3.1 MSE_*^h of Bootstrap Sample Mean: Exp(1).

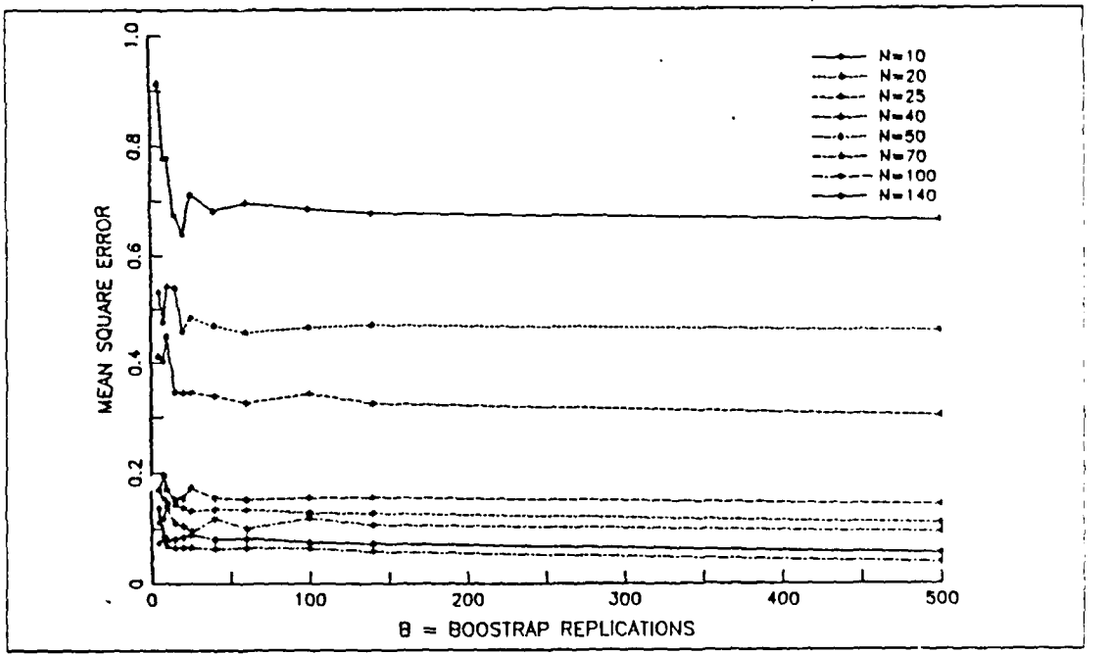


Figure 3.2 MSE_*^h of Bootstrap Sample Variance: Exp(1).

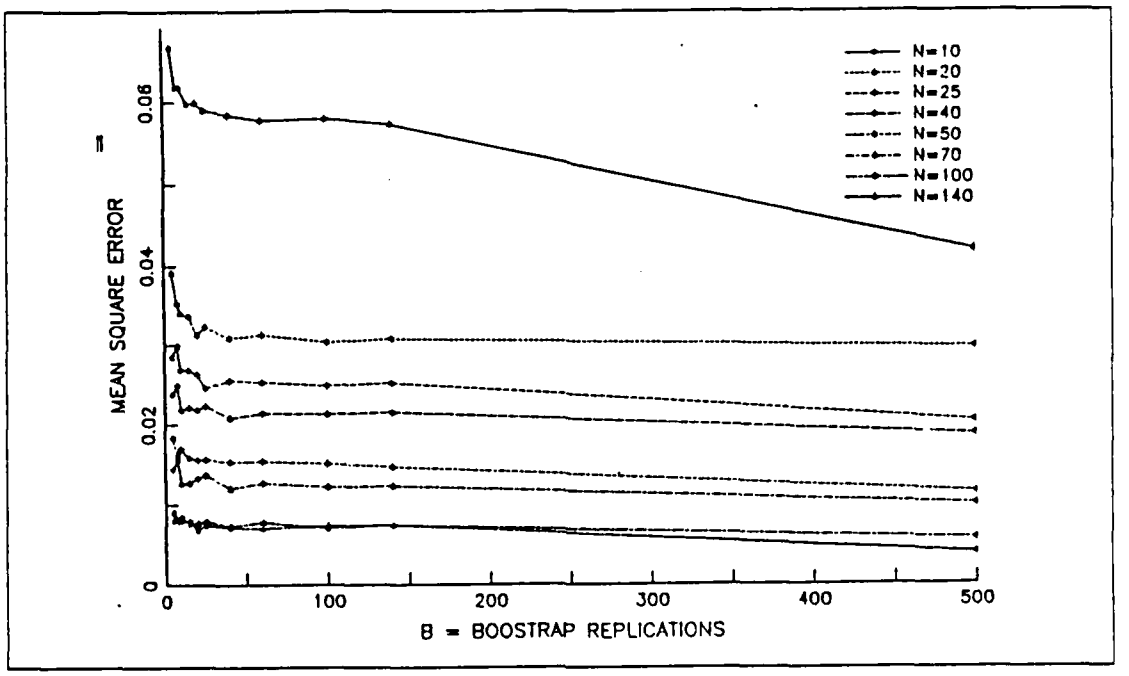
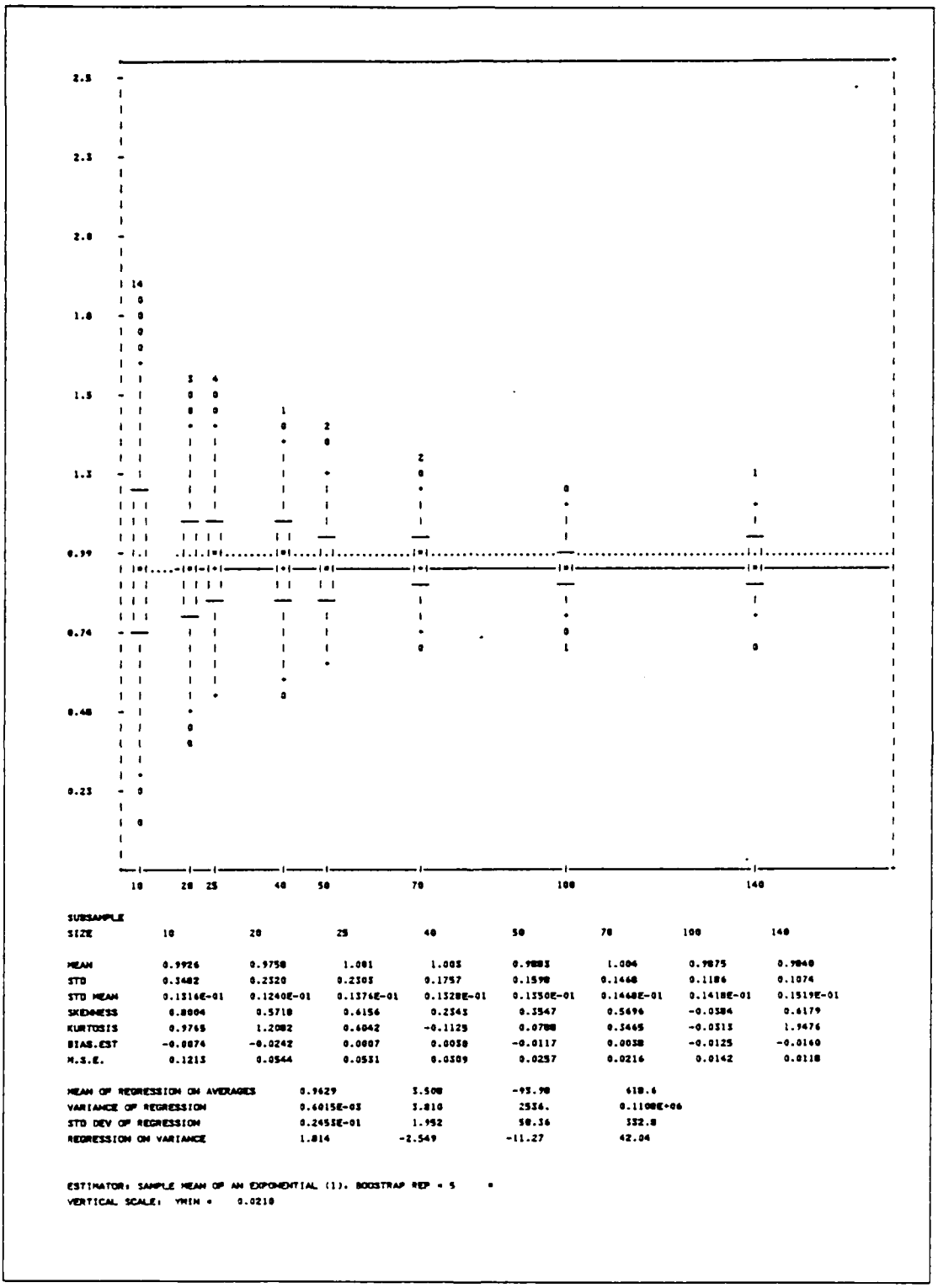


Figure 3.3 MSE_*^h of Bootstrap Coeff. of Variation: Exp(1).

beyond $B \geq 50$ is barely noticeable. For example, see Figure 3.2, the MSE_*^h of the sample variance decreases only by one-thousandth of a unit when B is increased from 200 to 500 replications. This is also true for the sample mean. However, for the coefficient of variation (see Figure 3.3), the MSE_*^h improved about two percent (.02) in the same range for a small sample size (n = 10). These results give an idea of the performance of the MSE of the bootstrap estimates of a given estimator. It should also suggest to the statistician that once the estimators are performing *fairly well* (i.e., once this threshold has been attained), there is no reason to increase the amount of bootstrap replications, since this will not induce a great improvement in the estimates. An important point here is that when an attempt is made to estimate the sample variance using the bootstrap method, the number of bootstrap replications should be greater than 100 in order to decrease the MSE_*^h below 0.6.

The bootstrap distribution of some of the estimators are shown in Figures 3.4, and 3.5 in the form of boxplots and a summary of the distributional statistics. These were obtained by using a statistical package, called SMTB10, developed at NPGS (See Appendix B). This package was modified by the author of this thesis in order to obtain MSE_*^h. Each boxplot represents the distribution of the bootstrap estimator based on the sample size n.



SUBSAMPLE SIZE	10	20	25	40	50	70	100	140
MEAN	0.9926	0.9758	1.001	1.003	0.9883	1.004	0.9875	0.9840
STD	0.3402	0.2320	0.2303	0.1757	0.1598	0.1468	0.1186	0.1074
STD MEAN	0.1316E-01	0.1240E-01	0.1376E-01	0.1328E-01	0.1350E-01	0.1468E-01	0.1418E-01	0.1519E-01
SKEWNESS	0.8804	0.5718	0.6156	0.2543	0.3547	0.5696	-0.0384	0.6179
KURTOSIS	0.9765	1.2082	0.6042	-0.1125	0.0788	0.3465	-0.0313	1.9476
BIAS.EST	-0.0074	-0.0242	0.0007	0.0038	-0.0117	0.0038	-0.0125	-0.0160
M.S.E.	0.1213	0.0544	0.0531	0.0309	0.0257	0.0216	0.0142	0.0118

MEAN OF REGRESSION ON AVERAGES	0.9629	5.508	-93.98	618.6
VARIANCE OF REGRESSION	0.6015E-03	3.810	2534.	0.1108E+06
STD DEV OF REGRESSION	0.2453E-01	1.952	50.36	332.8
REGRESSION ON VARIANCE	1.814	-2.549	-11.27	42.04

ESTIMATOR: SAMPLE MEAN OF AN EXPONENTIAL (1). BOOSTRAP REP = 5 *
VERTICAL SCALE: YMIN = 0.0210

Figure 3.4 Bootstrap Dist. of Sample Mean B = 5.

Notice, in Figure 3.4, that the distribution of the bootstrap sample mean resembles a Normal, as would be expected by the Central Limit Theorem, with the Kurtosis and Skewness oscillating around zero, as n increases. Recall from previous section that the standard deviation of X^*, in the case of Figure 3.4, would be estimated by

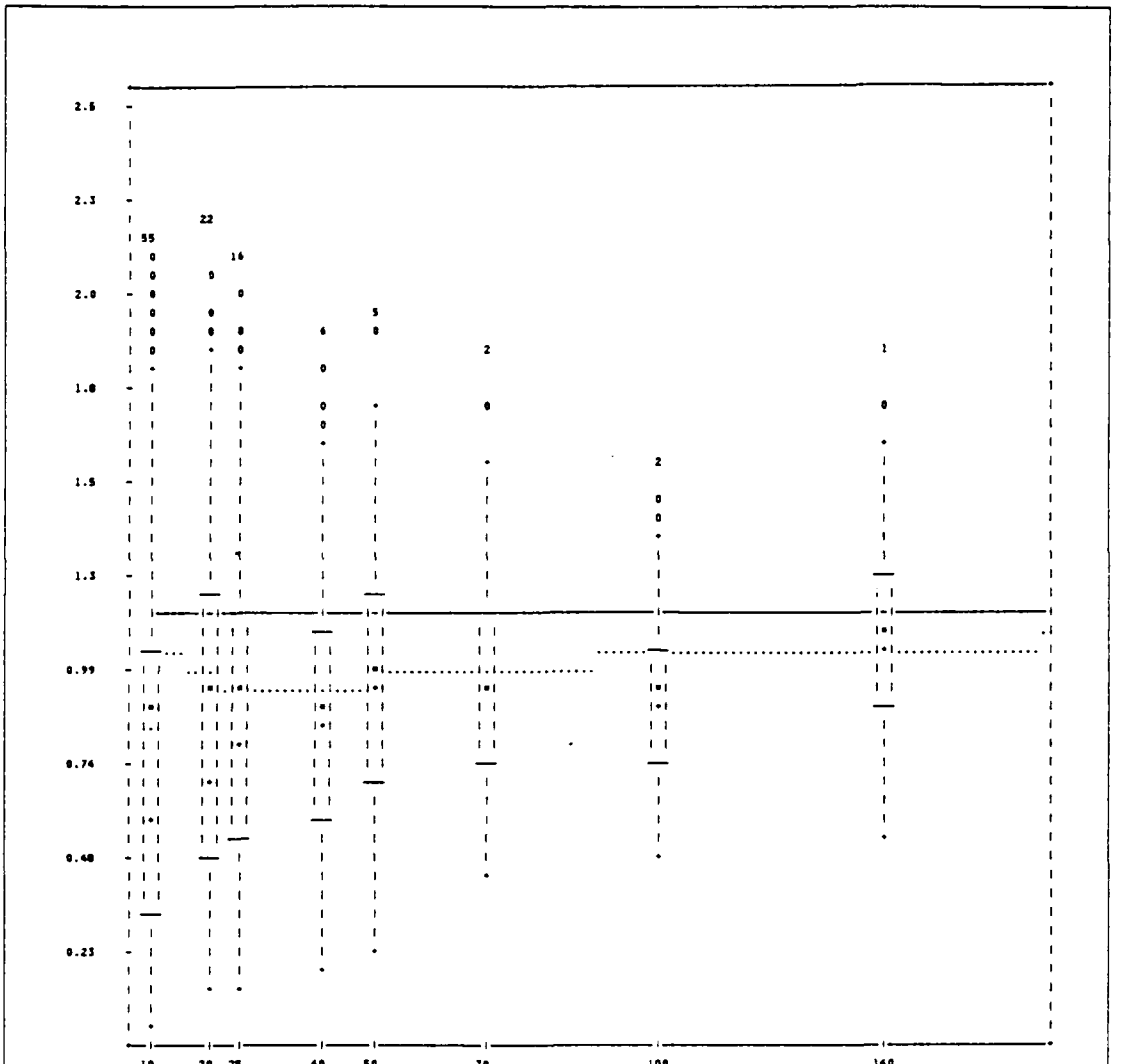
$$STD_*^h(X^*) = STD_*^h/\sqrt{n^*}, \qquad n^* = N \times M/NE(I)$$

and STD_*^h is the value shown on the bottom table of this figure. Figure 3.5 shows the distribution of the bootstrap sample variance (3.5). Looking at the distribution summary, one can say that this distribution is quite similar to that of a scaled Gamma(k,β) distribution. Again as n, increases the Kurtosis and Skewness get closer to that of the Gamma, say 6/k, and $2/\sqrt{k}$ respectively. Figure B.4 and B.5, Appendix B, show the distribution of the same estimators when B = 150. It is easy to see that the distributional characteristics for the estimators follow the same patterns as those discussed above, where B = 5. The only difference there is that, as expected, the number of outliers decreases significantly particularly in the case of the sample variance.

B. THE SAMPLE VARIANCE

This experiment was intended to further study the behavior of the bootstrap sample variance for populations with various distributions. The ones discussed in this section are the GAMMA(0.5,1), NORMAL(0,1) and LAPLACE(0,1). For this experiment, the sample size where n = 5, 10, 20, 25, 30, 50, 60, and B= 5, 8, 10, 15, 20, 25, 30, 35, 40, 50, 100, and 500. In the first two cases, the GAMMA and NORMAL distributions, the bootstrap sample variance seems to approximate the population variance fairly well when $n \geq 50$, where the MSE_*^h is less than 0.10. Figures 3.6, 3.7, and 3.8 show the relation between B, n, and the MSE_*^h of the bootstrap sample variance for a Gamma(0.5,1), Normal(0,1), and Laplace(0.1) respectively.

Notice that there is a lot of random variation in the MSE_*^h when B is in the range $5 \leq B < 50$ for $n \leq 30$, and for $B < 25$ when $30 < n \leq 60$. This random noise extends beyond these ranges in the case of the Gamma(0.5,1). Notice that in Figure 3.6, the lines for the MSE_*^h of the sample variance when n= 15, and 20 are above that when n= 10 for $B < 300$. However, when B= 500, these lines lie below the one corresponding to n= 10. The MSE_*^h for n= 15, and 20 is actually less than the MSE_*^h



SUBSAMPLE SIZE	10	20	25	40	50	70	100	140
MEAN	0.8992	0.9690	0.9856	0.9174	1.025	0.9887	0.9559	1.114
STD	0.9502	0.7282	0.6412	0.4628	0.412.	0.3544	0.2695	0.3512
STD MEAN	0.3591E-01	0.3892E-01	0.3832E-01	0.3045E-01	0.3483E-01	0.3344E-01	0.3218E-01	0.4967E-01
SKEWNESS	5.6228	2.1907	1.9856	1.1045	1.0087	0.7814	1.1468	1.4695
KURTOSIS	24.2171	6.1358	5.0364	1.1313	1.820'	0.9102	2.6711	4.3725
BIAS.EST	-0.1008	-0.0310	-0.0144	-0.0826	0.0230	-0.0115	-0.0441	0.1138
M.S.E.	0.9138	0.5315	0.4114	0.1696	0.1705	0.1120	0.0745	0.1363

MEAN OF REGRESSION ON AVERAGES	1.149	-15.07	345.8	-2202.
VARIANCE OF REGRESSION	0.1322E-02	15.05	9066.	0.3695E+06
STD DEV OF REGRESSION	0.3635E-01	3.879	95.19	607.8
REGRESSION ON VARIANCE	57.05	-827.3	4842.	-6976.

ESTIMATOR: SAMPLE VARIANCE OF AN EXPONENTIAL(1) BOOSTRAP REP = 5

Figure 3.5 Bootstrap Dist. of Sample Variance $B = 5$.

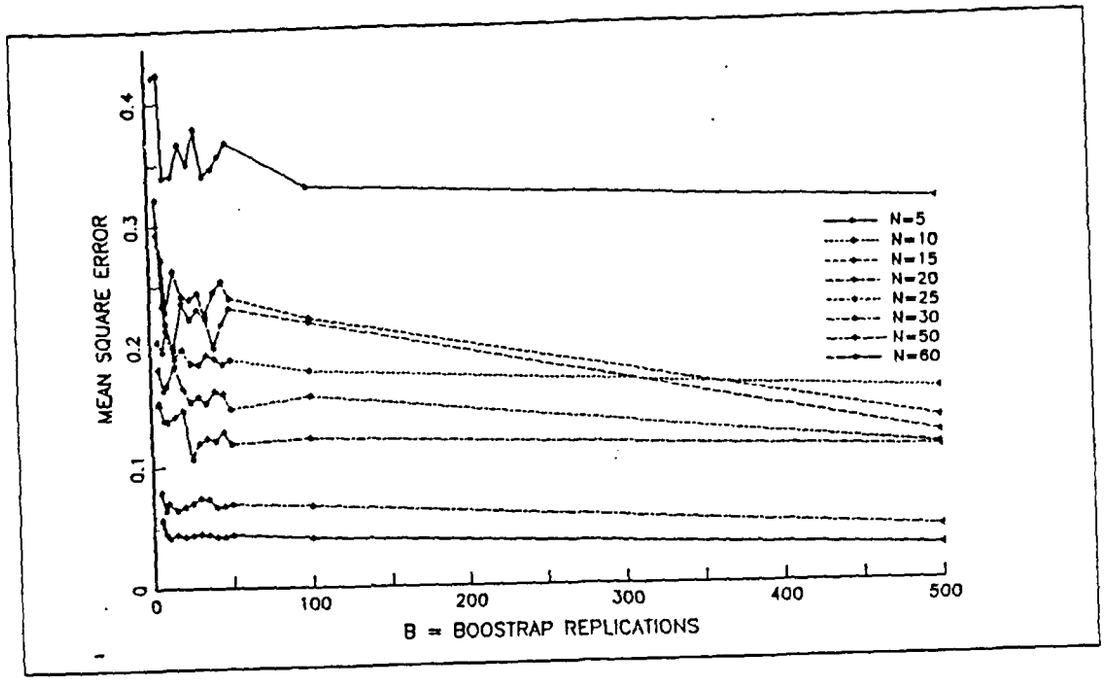


Figure 3.6 MSE_*^h of Bootstrap Sample Variance of a G(0.5,1).

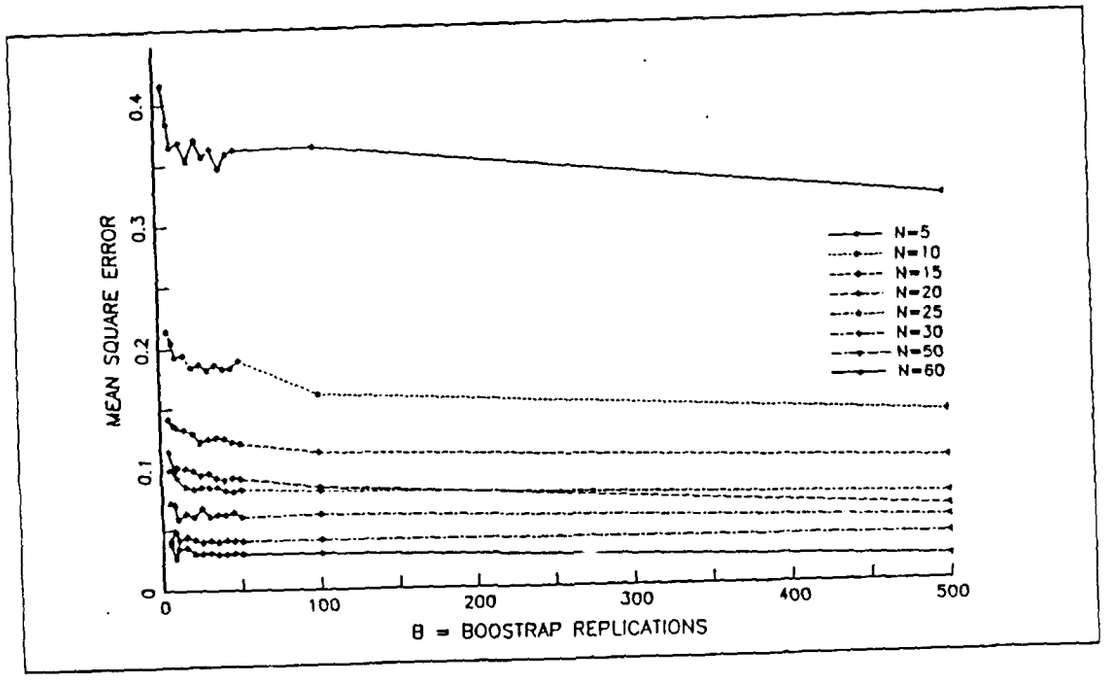


Figure 3.7 MSE_*^h of Bootstrap Sample Variance of a N(0,1).

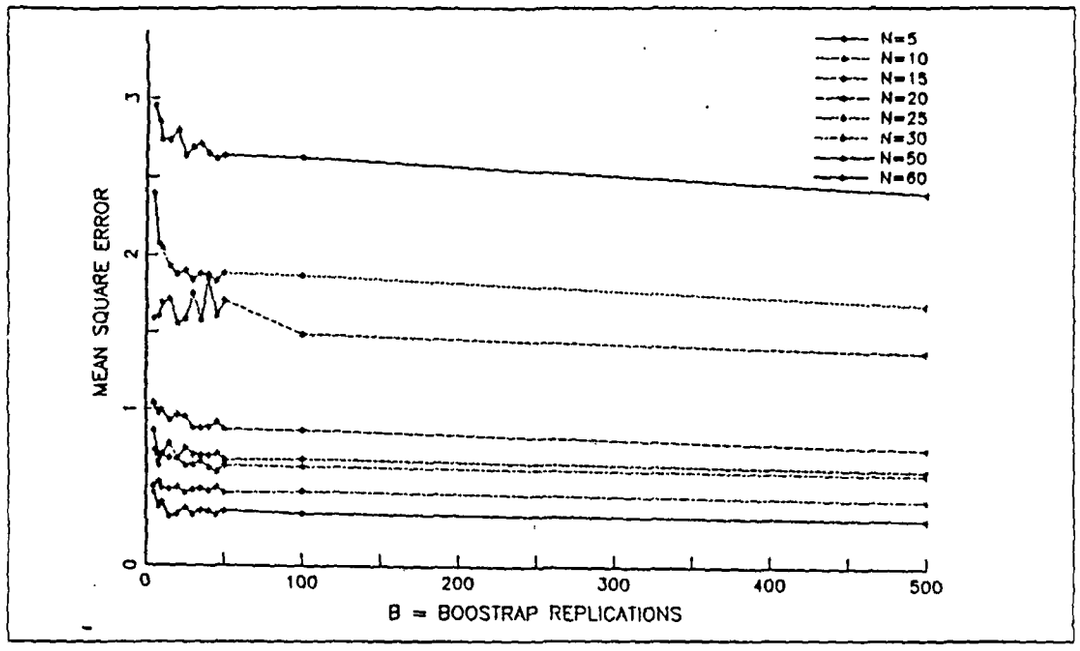


Figure 3.8 MSE_*^h of Bootstrap Sample Variance of a L(0,1).

for n= 10 just after B > 150. In this experiment, it is also true as found for the Exponential(1), that MSE_*^h decreases faster as n decreases than when B increases. This was also the result in the case of the Laplace(0,1). However (notice the scale of the MSE in this case), the MSE_*^h is quite high. Figure 3.8 shows that for a sample of size $n \leq 15$, the $MSE_*^h > 1.0$ even when B is as large as 500. It was suspected that probably this high MSE_*^h was caused by the mechanism used to generate Laplace random variates. The first method used in this experiment takes the difference of two Exponential (1) variates. The second method generates an Exponential(1) and converts it to a Negative-Exponential(1) with probability .5 . The histograms, using different sample sizes, showed that the first algorithm used to generate Laplace random variates was the most effective. In any case, the point here is that for the ranges of n and B used in the experiment, the MSE_*^h of the sample variance for a Laplace(0,1) never decreased below 0.2. This was not the case for the other distributions. This suggests that the performance of the bootstrap method depends on the distributional properties of the population in question as well as the estimator under consideration.

C. THREE DIFFERENT ESTIMATORS FOR THE VARIANCE

In Chapter Two, the expected value and the variance of the bootstrap sample mean (X^*) were derived. In this section, the expected value of the bootstrap sample variance, call this ${}_1S^{*2}$, is calculated. Let

$$
\begin{aligned}
{}_1S^{*2} &= [\textstyle\sum_i (X^*_i - \bar{X}^*)^2]/ (n - 1) \qquad (3.4)\\
&= [\textstyle\sum_i X_i^{*2} - n\bar{X}^{*2}] / (n - 1) \ .
\end{aligned}
$$

Note that

$$
E_*(X^*_i{}^2) = (1/n)\textstyle\sum_i X^2_i \qquad (3.5)
$$

so that

$$
E_*(\textstyle\sum_i X_i^{*2}) = \sum_i X^2_i \qquad (3.6)
$$

Likewise the second moment of $\bar{X}^*$ is given by:

$$
E_*(\bar{X}^{*2}) = (1/n^2)[\textstyle\sum_i X_i^{*2} + \sum_i\sum_j E(X^*_i X^*_j)] \qquad i \neq j \qquad (3.7)
$$

As before, $(X^*_j X^*_j)$ has probability $(1/n^2)$ of being any point of the form $(x_k x_l)$ so from (2.7)

$$
\begin{aligned}
E_*(X^*_i, X^*_j) &= (1/n^2)\, E[\textstyle\sum_i X_i^2 + \sum_i\sum_j X_i X_j \,] \qquad (3.8)\\
&= (1/n^2)\textstyle\sum_i X_i^2 + \sum_i\sum_j (X_i X_j)/n^2 \ .
\end{aligned}
$$

Now

$$
\begin{aligned}
\textstyle\sum\sum X^*_i X^*_j &= (n(n-1)/n^2)[\textstyle\sum_i X^2_i + \sum_i\sum_j X_i X_j] \qquad (3.9)\\
&= ((n-1)/n^2)(\textstyle\sum_i X_i)^2\\
&= n(n-1)X^2
\end{aligned}
$$

Then (3.7) can be expressed as

$$
E_*(\bar{X}^{*2}) = (1/n^2)[\textstyle\sum_i X^2_i + n(n-1)\bar{X}^2] \qquad (3.10)
$$

Finally, using (3.6) and (3.9), the conditional expected value of ${}_1S^{*2}$ is

$$E_*({}_1S^{*2}) = (1/(n-1))E_*(\sum_i X^*{}_i^2 + n\bar{X}^{*2}) \tag{3.11}$$

$$= 1/(n-1)[\sum_i E_*(X^*{}_i^2) - nE_*(\bar{X}^{*2})]$$
$$= 1/(n-1)\{\sum_i X_i^2 - [(1/n)(\sum_i X_i^2 + n((n-1))\bar{X}^2]\}$$
$$= 1/(n-1)[((n-1)/n)\sum_i X_i^2 - (n-1)\bar{X}^2]$$
$$= \sum_i (X_i^2 - \bar{X})^2 / n.$$

Call this $\sigma_S{}^{*2}$. Now suppose it is known that $X \sim N(\mu,\sigma^2)$- this restriction is not really required in this context - and it is desired to estimate the variance of X using the bootstrap method. As shown in the previous chapter,

$$E(\bar{X}^*) = \mu_X \ , \tag{3.12}$$

so the unconditional expected value of ${}_1S^{*2}$ is:

$$E({}_1S^{*2}) = E_*[E({}_1S^{*2}|X)] \tag{3.13}$$
$$= E[(\sum(X_i - X)^2)/ n]$$
$$= ((n-1)/n)\sigma_X{}^2$$

Then ${}_1S^{*2}$ is a biased estimator for $\sigma_X{}^2$. The finite population correction factor might thus be suggested to improve the performance of ${}_1S^{*2}$. Define

$${}_2S^{*2} = (n/(n-1))\,{}_1S^{*2} = n/(n-1)^2 \sum_i (X_i^* - \bar{X}^*)^2 \tag{3.14}$$

an unbiased bootstrap estimator of $\sigma_X{}^2$. Analyzing expression (2.5) and (3.11), yet another estimator for $\sigma_X{}^2$ can be suggested. Since the value of $E_*(\bar{X}_i^*) = \bar{X}$ is known, the following estimator for $\sigma_X{}^2$ also seems reasonable:

$${}_3S^{*2} = \sum (X^*{}_i - \bar{X})^2 / n \tag{3.15}$$

The third experiment was conducted to compare the performance of these three estimators (3.4), (3.14), and (3.15). Figures 3.9, 3.10, and 3.11 show the results of this experiment.

As can be seen, the third estimator, ${}_3S^{*2}$, in almost all cases outperforms the other two for all different sample sizes tried in this experiment. Even the second estimator (3.14) performs almost as good as ${}_1S^{*2}$ when $n > 50$. When $n \geq 50$, the

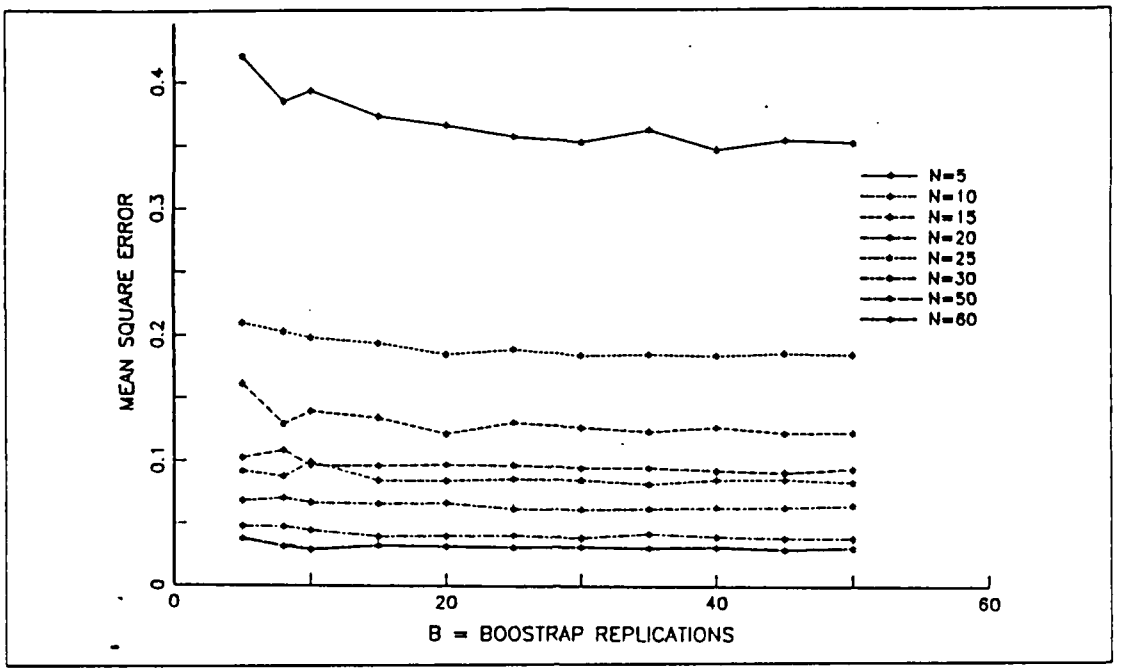


Figure 3.9 MSE_*^h of the Sample Variance of a N(0,1).

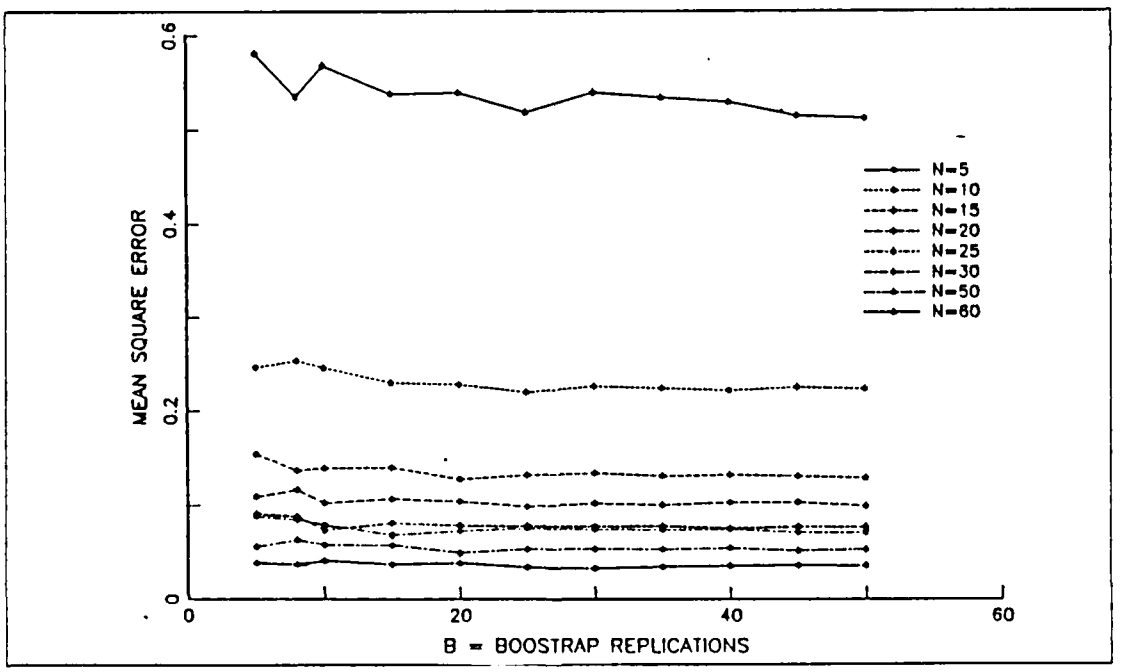


Figure 3.10 MSE_*^h of the 2nd Variance Estimator of a N(0,1).

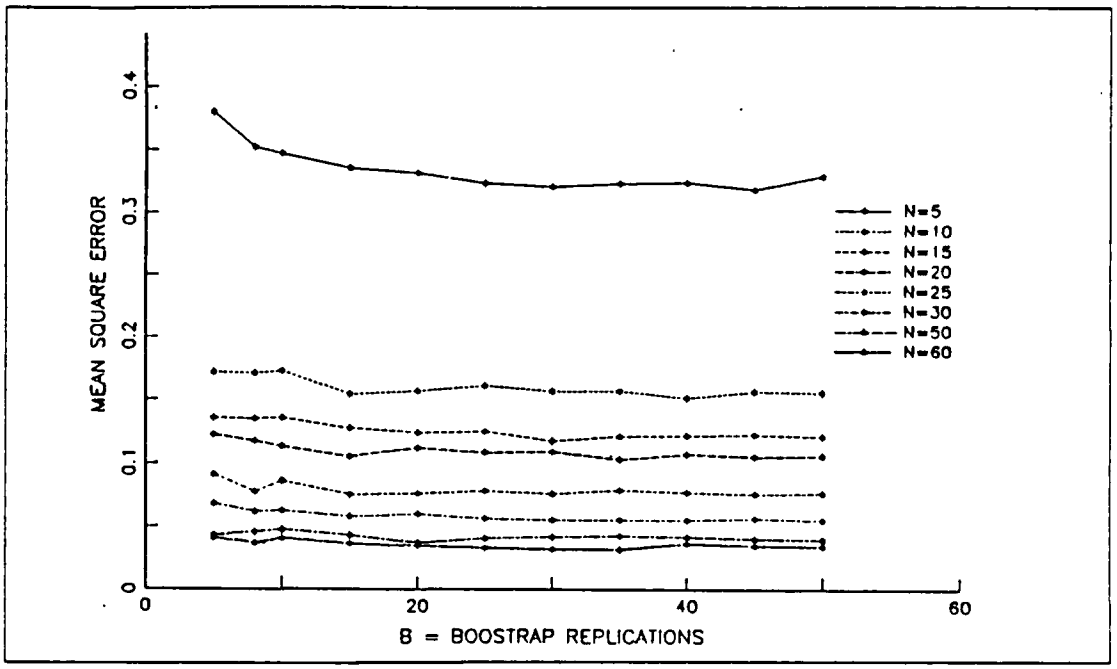


Figure 3.11 MSE_*^h of the 3rd Variance Estimator of a N(0,1).

difference between these three different estimators is barely noticeable. However, for very small samples, n < 20, ${}_3S^{*2}$ is definitly a better estimator for σ^2 than ${}_1S^{*2}$. Efron [Ref. 1] has suggested the use of ${}_1S^{*2}$ as the bootstrap estimator of the sample variance. As the plots suggest, it could be now recommended the use of ${}_3S^{*2}$ and even ${}_2S^{*2}$ (for larger samples, n > 50) rather than ${}_1S^{*2}$ to estimate the sample variance. Note that as $n \to \infty$, ${}_1S^{*2}$ is the same as ${}_2S^{*2}$. (Note: these two estimators (3.14) and (3.15) are called VARIA2 and VARIA3 respectively in the FORTRAN code, listed in Appendix A).

D. THE CENTER OF A DISTRIBUTION: COMPARISON OF THE MEAN, MEDIAN AND TRIMMED MEAN

The sample mean is the most used estimator for the center of a distribution. However, two other estimators are also used, specially for symmetric distributions: the median and the 5% trimmed mean. There have been many comparisons of the asymptotic performance of these three estimators. Lehman [Ref. 8] has calculated the asymptotic values of these estimators in case when the sample is from a Normal(0,1) or a Laplace(0,1) population. These calculations are summarized in Table 1 below.

TABLE 1

ASYMPTOTIC VARIANCE OF THE MEAN, MEDIAN AND 5% TRIMMED MEAN

Probability Distribution	ESTIMATOR		
	Mean	Median	5% Trimmed Mean
Normal(0,1)	1.0/n	1.57/n	1.01/n
Laplace(0,1)	2.0/n	1.00/n	1.65/n

These values, among other things, show that for the case of sample coming from a Normal(0,1), the mean has less asymptotic variance than the other estimators. However, if the data comes from a population with heavy tails, like the Laplace, the median is a better estimator asymptotically (having less variance). The 5% trimmed mean is a compromise between the other two: it should used when the practitioner does not know the nature of the tails of the population.

A fourth experiment was conducted to see if these observations hold when the corresponding bootstrap estimators are used. In this experiment, the MSE of of the bootstrap estimators were compared with the asymptotic MSE for the usual estimators as B increases. The asymptotic MSE (call it MSE_A) of the three estimators could be estimated by adding the asymptotic variance, as defined in Table 1, plus the bias-squared. The MSE_A was compared with the MSE_*^h of the bootstrap estimators, for several sample sizes, as B increases.

Figures 3.12, 3.13, and 3.14 summarize the results of this comparison for the case of a Normal(0,1) population. Figures 3.15, 3.16, and 3.17 show the results for a Laplace(0,1) population.

In these figures, the solid horizontal lines represent the values of the asymptotic MSE of the usual estimators. For example, in Figure 3.12 the estimated asymptotic MSE of the sample mean for a sample of size $n=5$ is approximately $1/5.0 + (BIAS)^2 \sim .20$. The dotted line represents the estimated MSE of the bootstraped estimators as B increases.

In summary, for the Normal(0,1) population, the bootstraped sample mean and the 5% trimmed mean have less error, asymptotically; they are estimating the center of

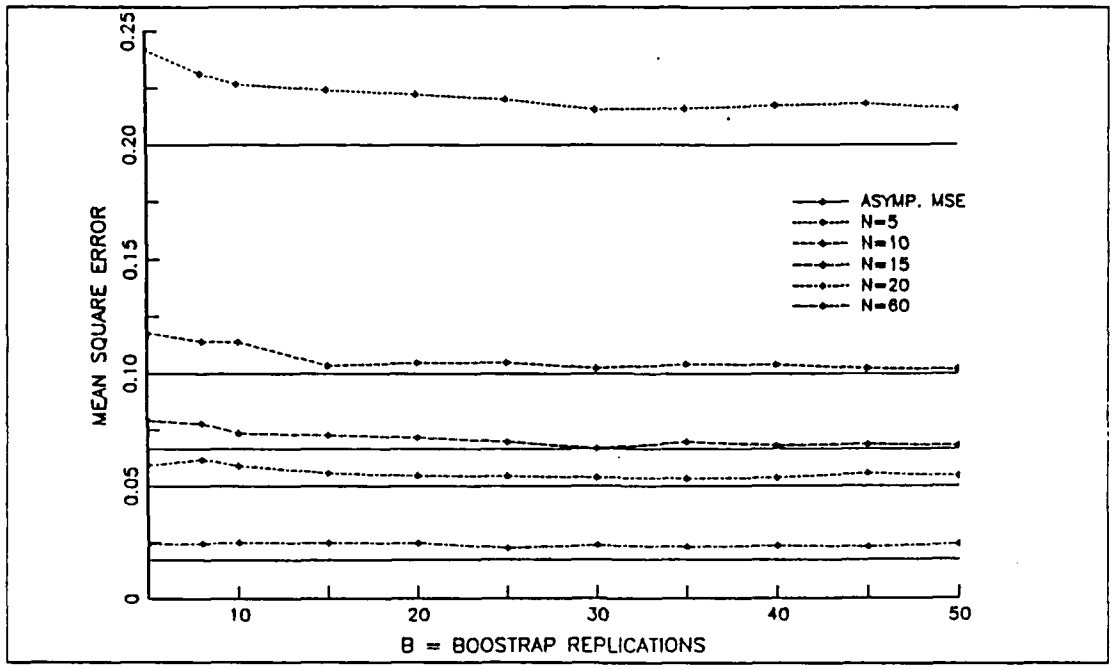


Figure 3.12 Asymptotic MSE of the Sample Mean of a N(0,1).

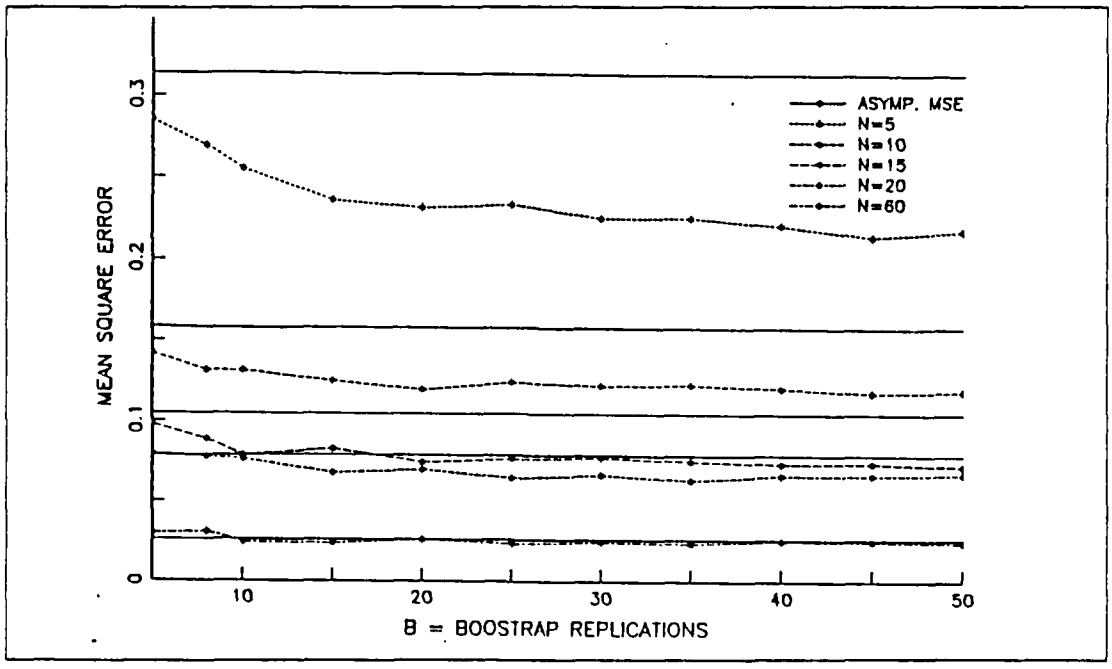


Figure 3.13 Asymptotic MSE of the Sample Median of a N(0,1).

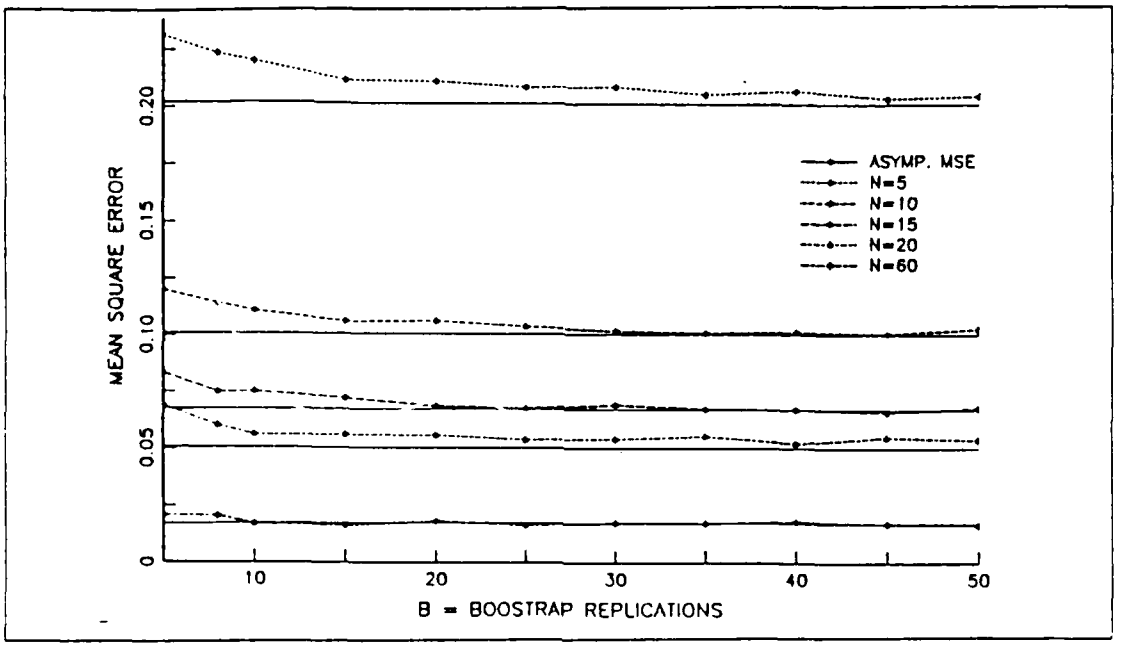


Figure 3.14 Asymptotic MSE of the Sample 5% Trimmed Mean of a N(0,1).

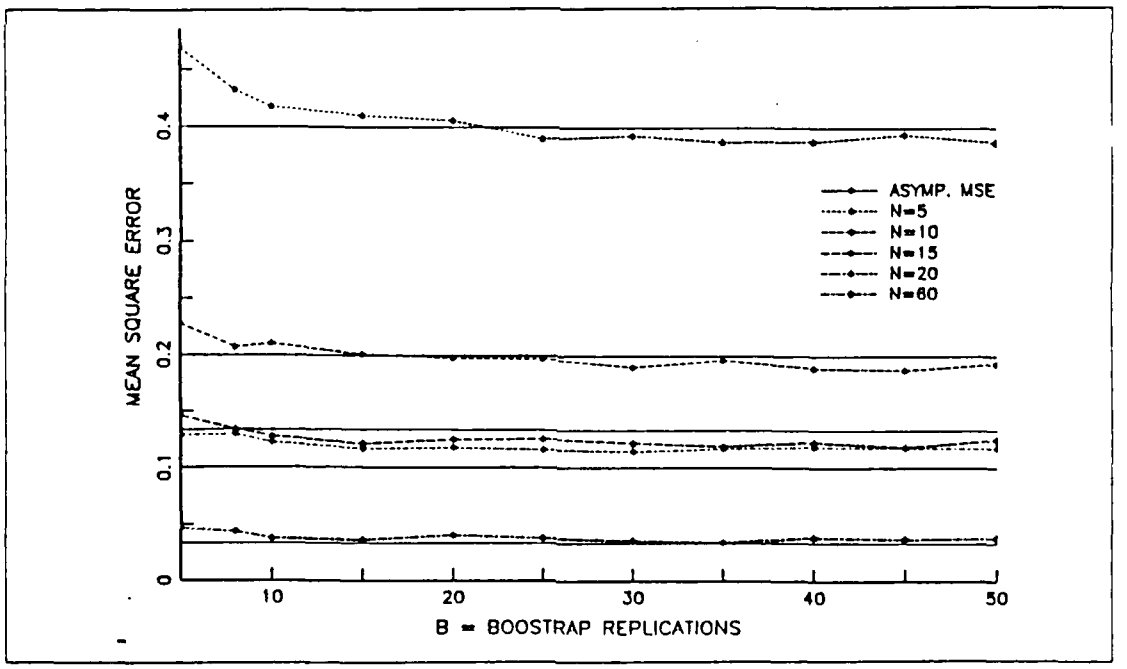


Figure 3.15 Asymptotic MSE of the Sample Mean of a L(0,1).

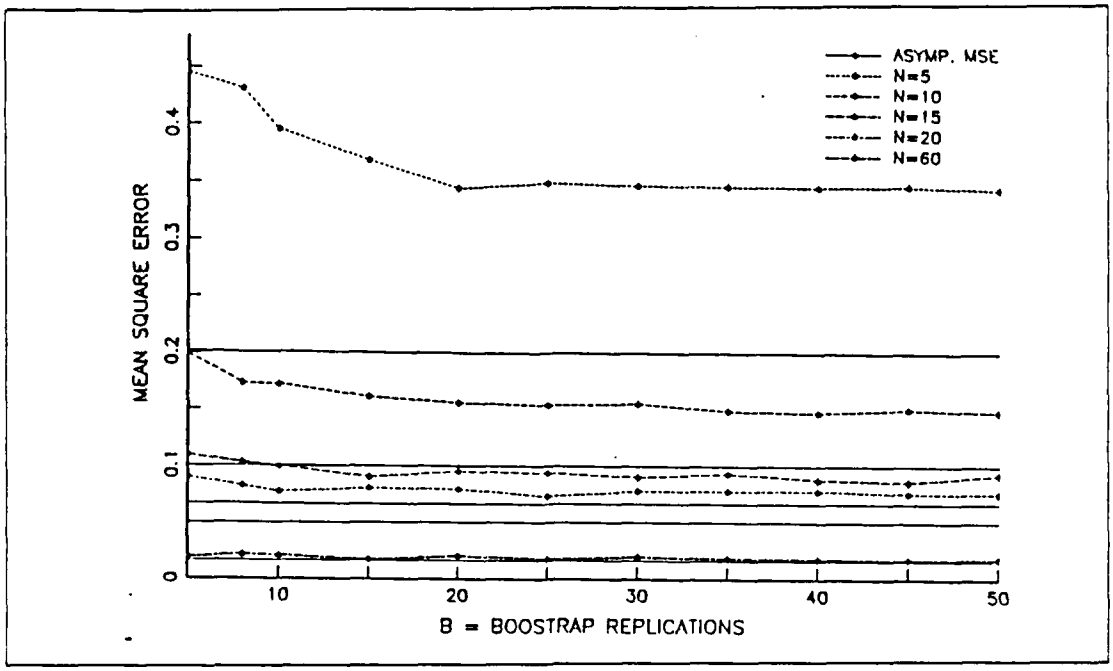


Figure 3.16 Asymptotic MSE of the Sample Median of a L(0,1).

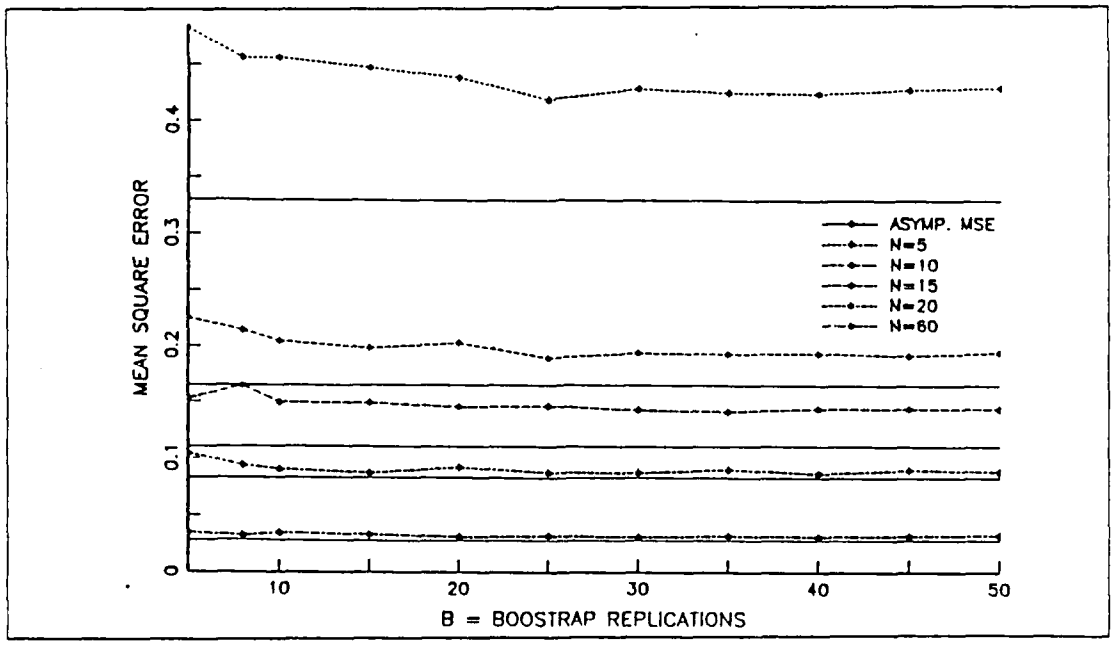


Figure 3.17 Asymptotic MSE of the Sample 5% Trimmed Mean of a L(0,1).

the distribution with much better precision than the bootstrap sample median. Comparing Figures 3.12 and 3.13, it looks obvious that for sample sizes $n \leq 60$ the bootstraped sample mean shows much smaller MSE than the bootstraped sample median. When the sample size is $n = 60$ there is no distinguishable difference between the estimated MSE's of these two estimators. Notice that the bootstraped 5% trimmed mean (Figure 3.14) seems to perform as well as the bootstraped sample mean; it is better for very small samples, say for $n = 5$, 10, and 15. This confirms the general relationship among these estimators, even in the case of bootstraping the estimators, that the 5% trimmed mean is a robust compromise between the sample mean and the sample median.

The results obtained in this experiment, however, do not agree with the classical theory in the case of the Laplace population. In this case the bootstraped sample mean outperforms the bootstraped sample median in estimating the center of the distribution, for sample size $n \leq 20$. For a sample of size $n = 60$, there is no real difference between these two estimators, in terms of $MSE_*{}^h$. Notice that the 5% trimmed mean (Figure 3.17) performs better than the bootstraped sample median (Figure 3.16) for the cases where $n < 60$, but in turn, is outperformed by the bootstraped sample mean (Figure 3.15).

E. LINEAR REGRESSION BY BOOTSTRAPING THE RESIDUALS

In a final experiment, linear regression estimation was considered. In this case, there is a choice of bootstraping methods; however, in this thesis only one method is considered. The method considered here relies on bootstraping residuals to estimate the variance of the β^h vector(β^h stands for " β hat"). A measure to estimate the MSE of this vector is also introduced.

In the typical linear regression problem there are n independent observations (real-valued) Y_i and it is assumed that the following model holds:

$$Y = X\beta + \varepsilon \ , \tag{3.16}$$

where ε is a random sample from some population F, and β is a $p \times 1$ vector of unknown parameters that must be estimated. All that is assumed about F is that it is centered at zero, $E(\varepsilon) = \mathbf{O}$ and $Cov(\varepsilon) = \sigma^2 I$. One way of estimating β is by the commonly used *least squares* method, in which the sum of the squared distances

(3) Using the same fitting technique used to obtain β^h in the original problem, calculate β^* . Then obtain an estimate of β^* :

$$\mathfrak{b}^* = (X'X)^{-1} X'Y^* \tag{3.21}$$

(4) Repeat steps (2) and (3) B times obtaining independent bootstrap realizations $\mathfrak{b}^*_1, \mathfrak{b}^*_2, \ldots, \mathfrak{b}^*_B$. Then the covariance of β^h can be estimated by the sample covariance matrix of the $\mathfrak{b}^*_b$, b= 1, 2,..., B.

Efron has shown (See [Ref. 1: page 18]) that as $B \rightarrow \infty$,

$$\mathrm{Var}(\beta^*) = ((n-p)/n)\,(X'X)^{-1}\sigma^2 \tag{3.22}$$

where σ^2 is an unbiased estimate of the variance of Y_i . In this procedure, σ^2 can be estimated by ${}_2S^{*2}$. It can be seen that as $B \rightarrow \infty$,

$$\mathrm{Var}(\beta^*) \rightarrow \mathrm{Var}(\beta^h)\,. \tag{3.23}$$

The following experiment was conducted to estimate the MSE of β^h. Suppose it is known that the observations Y_i come from a Normal(0,1). Then the true value of the β- vector in the regression model (3.17) is β = (0,0,0), so the $E(\beta) = \mathbf{O}$ and the variance-covariance matrix of β is $\Sigma_\beta = \sigma^2(X'X)^{-1}$, where it is known that $\sigma^2 = 1$.

For this experiment, a design matrix X of orthogonal-column vectors was created. This matrix has 1's in the first column; then a series of n alternating 1's and -1's in the second column; and finally the third column (for p= 3) is a series of two 1's and two -1's (also, $n = 2^x$, x = 2, 3, 4,...). Then it was possible to readily calculate β^h , by

$$\beta^h = (1/n)\,(X'Y)\,. \tag{3.24}$$

The bootstrap algorithm described above was used to generate a sample of β^*_i . Then, an estimate of β_i^* is

$$\mathfrak{b}^*_i = (1/n)\,(X'Y^*)\,. \tag{3.25}$$

It was desired to develop a measure of precision for β^* analogous to MSE, which depends on $Var(\beta^*)$ and the bias of β^* . Define

$$MSE(\beta^*) = E[(\beta^* - E(\beta^*))^2] . \qquad (3.26)$$

Recall that in this experiment the $E(\beta^h) = \mathbf{O}$. Then, (3.26) could be estimated in the following way:

1) Do step (4), as above, obtaining

$$MSE_*(\beta^*) = [\sum_i (\beta_i^* - E(\beta^h))^2] / B \qquad i = 1, 2, \ldots, B \qquad (3.27)$$
$$= [\sum_i \| \beta_i^* - \beta \|^2] / B .$$

2) Repeat (1) a number of M times to obtain an average $MSE_*{}^h$ of the procedure (3.27).

The results of this experiment are shown in Figure 3.18.

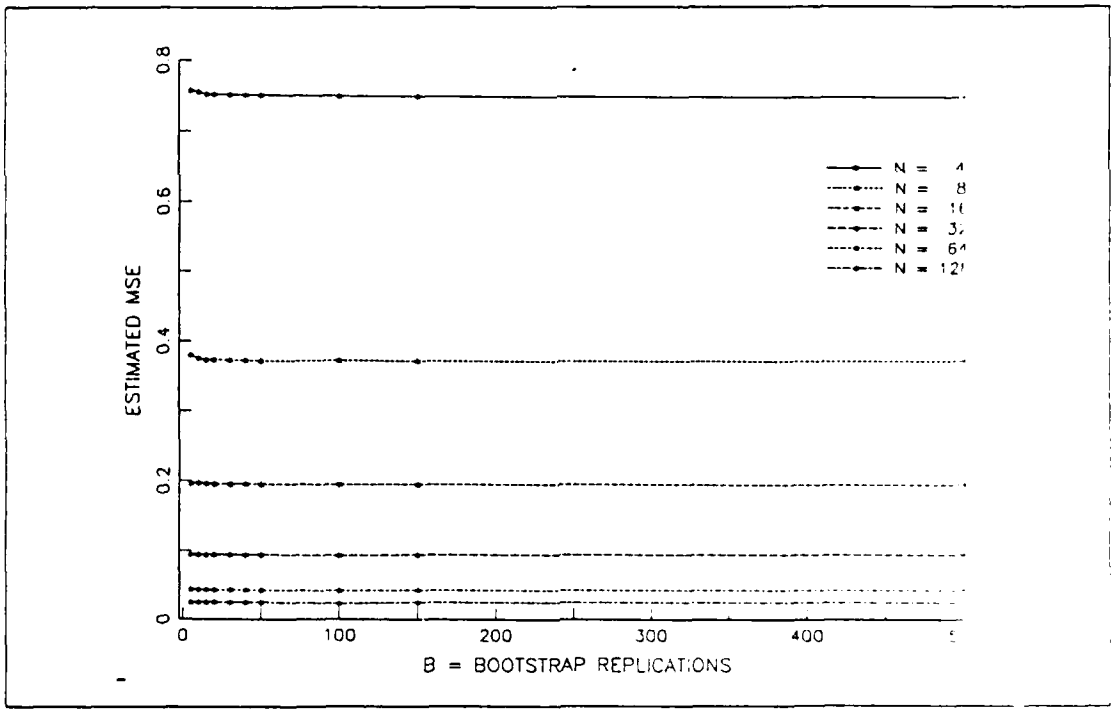


Figure 3.18 Estimated Averages MSE of β^h.

Here, the sample sizes were taken as n = 4, 8, 16, 32, 64, and 128, and M = 15. The estimator β^* was bootstraped a number B = 5, 10, 15, 20, 30, 40, 50, 100, 150, and 500. The results obtained were surprising. When the number of observations is small, $n < 33$, the $MSE_*{}^h$ of the estimator is relatively high ($MSE_*{}^h > .09$) even when B is as large as 500. When $n > 65$, there is some improvement in the $MSE_*{}^h$; in this case, the $MSE_*{}^h$ is at least 5% lower that when to $n < 33$. It is interesting to see that increasing B from 5 to 500 there is no remarkable gain in the precision of estimator when $n > 65$; the $MSE_*{}^h$ oscillates around the same value. Now, when $n < 33$, increasing B by the same amount, the $MSE_*{}^h$ decreases but less than 1% of its initial value. It seems that in the linear regression estimation the key problem is the size of n and not of B.

When using this method for estimating the MSE of β^h , the practitioner must bear in mind that it involves the residual distribution and hence assumes that the linear model is correct.

IV. CONCLUSIONS

As it has been shown, the Bootstrap is an accurate method for estimating the precision of the estimates and for estimating the distribution (or some feature of the distribution) of an estimator. For MSE, the number B required to obtain a certain degree of accuracy will vary depending mainly on the population (this is a subject for further studies) and the type of the estimator used for estimation. It was found that when the sample comes from a population having heavy-long tails, such as the Laplace distribution, the bootstrap estimator for the mean is a better estimator for estimating the center of the distribution than the median or the 5% trimmed mean; where in the case of using nonbootstrap estimators, the median is a better estimator than the other two estimators.

In estimating the variance of a population, it was found that there exists an estimator that is more accurate than the typical estimator recommended in the bootstrap literature. This estimator (${}_3S^{*2}$) relies on the fact that the original sample mean in the bootstrap method is known. Once this value is calculated, there is no need to find $\bar{X}^*$ for each bootstrap sample, since $\bar{X}$ is fixed through the process. Another estimator for σ^2 was also proposed, ${}_2S^{*2}$. This estimator is unbiased, where ${}_1S^{*2}$ is not, but for small sample sizes, $n < 30$, is not as accurate as ${}_3S^{*2}$. It should be emphasized that in using this estimator, ${}_3S^{*2}$, one can reduce the computer time required to estimate σ^2. Hence, this is another advantage in using this estimator.

In the linear regression estimation, using as a measure of precision definition (3.28), it was found that the bootstrap method analyzed in this thesis gives estimates with small $MSE_*{}^h$ with relative small sizes of B, but for relatively large sample size, $n > 60$. When the sample size is small, increasing B up to 500 will result in a gain of around 1% in the precision of the estimates. Thus, in the linear regression estimation the critical issue for MSE is the sample size. It was also noted that the disadvantage of this method is that it assumes that the model in question is correct.

The result that seems to apply to all cases studied in this work is that, in using the bootstrap method for estimating MSE of some parameter θ , there really exits a tradeoff between B and n: as n increases, one can significantly decrease B and still get very precise estimates. However, no matter what n is, once some degree of accuracy

has been obtained, there is no reason to increase B much more since this will not induce greater precision in the estimates. In Appendix C , the reader will find tables that provide information about this tradeoff for given estimators and populations. Analyzing the figures presented in previous chapters and these tables, a rule of thumb about the relation between n and B can be hypothesized. The following rule seems reasonable: make the number $B \sim 1000/n$. In almost all cases studied here, this rule yielded estimates with $MSE_*^h < 0.05$ (note: independent of n, making $40 < B > 60$ will also produces estimates with small MSE_*^h). The only exception is when the population in question was Laplace(0,1). This is an area that needs further study.

Finally, it was found that a (possibly not serious) disadvantage in using the bootstrap method is the computer time required to obtain the estimates. For example, in estimating the variance of a Gamma(0.5,1) distribution, increasing B from 20 to 100 increased the CPU time of the IBM 3033-A16 system used in this experiment about 75%. This time is increased at least another 50% if one desires to obtain the distributional characteristics of the estimator (i.e., boxplots). However, in view of the decreasing cost of computer time, this does not seem to be a major obstacle for using this method.

APPENDIX A

LIST OF SPECIAL NOTATIONS

(1) θ^h :θ -hat, estimator of θ

(2) F^h :empirical probability distribution

(3) $\theta^*(F^*)$:the value of θ based on bootstrap method

(4) X^* :a bootstrap random sample

(5) $MSE_*{}^h$:estimated MSE based on bootstrap method

(6) β^h :estimator of the $p \times 1$ β -vector

(7) $\mathfrak{b}^h$:an estimate of β^h

(8) β^* :estimator of β based on bootstrap method

(9) $\mathfrak{b}^*$:an estimate of β^*

APPENDIX B
FORTRAN CODE FOR BOOTSTRAPING

This program, called BOOTST, was developed to estimate distributional properties of some statistical estimators using the Bootstrap Method. Also it is possible to obtain estimates of the MSE of the estimators. The code was written in FORTRAN 77. It can generate a random sample for Monte Carlo simulation or can read the sample data by a CALL to a subroutine FDATA (at the end of the code listed below). The user can generate samples from the following distributions: Exponential(λ), Laplace(0,1), Uniform(0,1), Normal(0,1), Gamma(α,1), Poisson(λ), and the Geometric(p). The parameters α, λ, and p can be specified by the user within the appropriate function. With this program, the user can study the distributional properties of the following bootstrap estimators: mean, variance coefficient of variation, serial correlation, median, and the 5%-trimmed mean. Also, one can obtain estimates of the "β -vector" in the case of the linear regression estimation by bootstraping the residuals (See Chapter Three, Section D). The program is structured in five main sections: the MAIN program, to include input requirements; the DATA GENERATION, the ESTIMATORS definition, the BOOTSTRAP SAMPLING mechanism, and the STATISTICS sections.

The program can be used in two ways. The first, makes use of another program called SMTB10. This code was developed at the NPGS by Prof. P.A.W. Lewis, and Mr. Luis Uribe (See [Ref. 9]). It is highly recommended that the user become familiar with the documentation of STMB10 before attempting to use BOOTST. In general, when using this option, the user must create an input file containing the parameters specified in the input section of BOOTST. Then, a CALL is made to STMB10, and in turn STMB10 will make various sequential calls to generate the data, calculate the values of the desire estimators (using the bootstrap mechanism), and produce the statistics. When a call to STMB10 is made, the user could produce estimates for 1, 2, or 3 different estimators using 1, 2, or 3 sample data generators or any of the eight possible combinations. Also, the user could select up to 8 different sample sizes for each estimator. Therefore, in one execution, statistics for up to three different estimators, using up to three different data generators, and for up to eight different

sample sizes can be obtained using the bootstrap method. These options are controlled in the INPUT requirements of BOOTST. At the end of each execution, BOOTST will send to a printer (or to the screen, depending on the option selected) a file containing boxplots and a summary of the statistics for each estimator. The input requirements are controlled by the user in a file called BOSIN.

The general execution of BOOTST runs as follows:

(1) *For each estimator*

(2) *Read Input Requirements* (*MAIN*)

(3) *CALL STMB10*

(4) *CALL Data Generator* (*Data Generation Section*)

(5) *$N = k \times n$ random variates are generated, where $k = 1$ or 2,..., or 8 different sample sizes. Then the data is sectioned into samples of sizes $N(K) = n$. If M repetitions of the process are allowed, then a total of $M \times N$ random numbers are obtained. Estimates are calculated for each sample size $N(K)$.*

(6) *CALL Estimator Function* (*Estimator Section*)
Begin Generation of Estimates

(7) *For I = 1 to B*
CALL BOOTSTRAP (*Bootstrap Section*)
CALL STATISTIC
Store Bootstrap Estimates
CALL STATISTIC
Store Mean of Bootstrap Estimates

(8) *PRODUCE Boxplot and Statistics*

The input requirements specific to BOOTST are explained below, the other inputs declared in the MAIN are specific to STMB10 (See [Ref. &ref10]).

(1) ANS : 1 or 0 : If the user wants to store each bootstrap estimate for each estimator, the answer should be 1. Estimates are stored in FILE 21.

(2) NE(I): a vector containing the sample sizes (n). Up to 8 different sample sizes.

(3) IB: Number of bootstrap replications for each execution.

(4) IX: Seeds used to generate data (up to 3 different seeds).

If the user desires to obtain estimates and graphical displays of two or more different estimators and is using a large number B, say $B \geq 60$, the amount of computer time required will increase significantly depending on the system used.

The second way to execute BOOTST is recommended for more experienced users or for those who do not want to obtain boxplots of the estimates. This option will save a great deal of CPU time. For this option, the user will have to make some simple changes to the MAIN program:

(1) Delete from the input requirement section those inputs that only apply to STMB10 (those not listed above).

(2) Replace the call to STMB10 by the following sequence of calls:

(i) Call Data Generator (i.e., one of the data generators)
(ii) Call Estimator (i.e., one of the estimator functions) The estimator function (subroutine) will make the appropriate call to the Bootstrap and Statistic subroutines.

(3) For this option, the input parameters ANS must be set to integer 1. Also, if the user now make reference to the code, it will be noticed that each estimator subroutine has a special parameter WI. This parameter must be deleted everywhere since its only applies to STMB10.

The computer code is listed below.

```
C
C     UPDATED      07-03-86    W. CORTES-COLON
C     MAIN : DECLARATION, INPUT SECTION AND CALL FOR SMTB10.
C
      COMMON IB,IX1,IX2,IX3,IX4,ANS
      COMMON Z(20000)
      CHARACTER*80 T1, T2, T3
      REAL*4  Y(10000),YMIN,YMAX, PMEAN(3),AMSEC(3)
      INTEGER  NE(8),D,RG,SEI,SVS,N,M,L,NEST,NSR
      INTEGER  IX1,IX2,IX3,IX4,IB,ANS
      EXTERNAL XMEAN,VARIA,COEVA,SECOR,MEDIA,TRIMM,VARI2,VARI3,BLREG
      EXTERNAL EXPON,UNIFO,NORML, GAMAF, POISF, GEOMF,LAPLA
C
C
      OPEN(UNIT=19, FILE='BOSIN')
      READ(19,*) ANS
   10 READ(19,*, END=999) N,M,L,D,RG,SEI,SVS,NEST,NSR
      READ(19,*) YMIN, YMAX
      READ(19,*) (NE(I),I=1,L)
      READ(19,*) IB
      WRITE(22,105) IB,(NE(I),I=1,L)
  105 FORMAT(I4,8I4)
      READ(19,*) IX1,IX2,IX3,IX4
      READ(19,115) T1
  115 FORMAT(A80)
      READ(19,115) T2
      READ(19,115) T3
      READ(19,*) (PMEAN(I),I=1,3)
      READ(19,*) (AMSEC(J),J=1,3)
C     CALL FOR SMTB10: PRODUCES BOX-PLOT AND COMPARISON OF STATISTICS
C
      CALL SMTB10(IX1,IX2,IX3,Y,N,M,NE,L,D,NSR,RG,SEI,SVS,YMIN,YMAX,
     *  NEST, NORML,XMEAN,T1,NORML,MEDIA,T2,NORML,TRIMM, T3,
     *  PMEAN,AMSEC)
      GO TO 10
  999 WRITE(6,*) 'END OF DATA INPUT'
      STOP
      END
C     DATA GENERATION SECTION
C
      SUBROUTINE EXPON(IX,X,NEK)
      REAL X(1)
      IF(NEK .LE. 0) RETURN
      CALL SEXPN(IX,X,NEK,1,0)
      RETURN
      END
C
      SUBROUTINE LAPLA(IX,X,NEK)
      INTEGER ISEED
      REAL X(1),XU(1000),X2(1000)
      IF(NEK.LE.0) RETURN
      CALL SEXPN(IX,X2,NEK,1,0)
      CALL SEXPN(IX,XU,NEK,1,0)
      DO 10 I=1,NEK
        X(I)=X2(I)-XU(I)
   10 CONTINUE
      RETURN
      END
C
      SUBROUTINE UNIFO(IX,X,NEK)
      REAL X(1)
      IF(NEK .LE. 0) RETURN
      CALL SRND(IX,X,NEK,1,0)
      RETURN
      END
C
      SUBROUTINE NORML(IX,X,NEK)
      REAL X(1)
      IF(NEK .LE. 0) RETURN
      CALL SNOR(IX,X,NEK,1,0)
      RETURN
      END
C
      SUBROUTINE GAMAF(IX,X,NEK)
      REAL X(1), ALPHA
      ALPHA=0.5
      IF(NEK .LE. 0) RETURN
      CALL SGAMA(IX,X,NEK,1,0,ALFA)
      RETURN
      END
C
      SUBROUTINE POISF(IX,X,NEK)
      REAL X(1),LAMDA
      LAMDA=0.5
      IF(NEK .LE. 0) RETURN
      CALL SPOIS(IX,X,NEK,1,0,LAMDA)
```

```
      RETURN
      END
C
      SUBROUTINE GEOMF(IX,X,NEK)
      REAL X(1), P
      P=0.5
      IF(NEK .LE. 0) RETURN
      CALL SGEOM(IX,X,NEK,1,0,P)
      RETURN
      END
C
CCC   ESTIMATOR SECTION : BRLG IS USED FOR LINEAR REGRESSION ESTIMATION
CCC   ONLY. IT IS RECOMMENDED TO USE THIS ESTIMATOR SEPARETLY: I.E,
CCC   WHEN CALLING SMTB10, USE ONLY ONE ESTIMATOR.
C
      REAL FUNCTION BLREG(YOBS,NEK,WI)
      COMMON IB,ANS
      REAL  YOBS(1),BMSTAR(3),MSEBS
      REAL  XDES1(600,3),XTRANS(3,600),XDES2(3,600),XTXINV(3,3)
      REAL  RES1(600),YHAT(600),RSTAR(600),BHAT(3),YSTAR(600)
      REAL  BSTAR(3)
      INTEGER  WI
      DO 10 I=1,NEK
         YHAT(I)=0.0
         DO 10 J=1,3
            XDES1(I,J) = 1.0
            XDES2(J,I) = 0.0
            XTRANS(J,I)=0.0
   10 CONTINUE
      DO 20 I=1,NEK,2
         XDES1(I,2)=-1.0
   20 CONTINUE
      DO 30 I=1,NEK,4
         XDES1(I,3) = -1.0
         XDES1(I+1,3) = -1.0
   30 CONTINUE
      DO 40 I=1,3
         XTXINV(I,I)=1.0/FLOAT(NEK)
         BHAT(I)=0.0
   40 CONTINUE
      DO 50 J=1,NEK
         DO 50 I=1,3
            XTRANS(I,J)=XDES1(J,I)
   50 CONTINUE
      DO 60 K=1,3
         DO 60 J=1,NEK
            DO 60 I=1,3
               XDES2(K,J)=XDES2(K,J) + XTXINV(K,I)*XTRANS(I,J)
   60 CONTINUE
      DO 70 K=1,3
         DO 70 J=1,NEK
            BHAT(K)=BHAT(K) + XDES2(K,J)*YOBS(J)
   70    CONTINUE
      DO 90 J=1,NEK
         DO 80 I=1,3
            YHAT(J)=YHAT(J) + XDES1(J,I)*BHAT(I)
   80    CONTINUE
         RES1(J)=YOBS(J)-YHAT(J)
   90 CONTINUE
      DO 95 IWX=1,3
         BMSTAR(IWX)=0.0
   95 CONTINUE
      MSEBS=0.0
      DO 100 IW=1,IB
            DO 110 JI=1,NEK
               RSTAR(JI)=RES1(JI)
  110       CONTINUE
            CALL BOOTS(RSTAR,NEK)
            DO 120 K=1,NEK
               YSTAR(K)=YHAT(K) + RSTAR(K)
  120       CONTINUE
            DO 130 K=1,3
               BSTAR(K)=0.0
               DO 130 KI=1,NEK
                  BSTAR(K)=BHAT(K) + XDES2(K,KI)*RSTAR(KI)
  130       CONTINUE
C           WRITE(6,5) (BSTAR(KL),KL=1,3)
C    5      FORMAT(3F8.4)
            DO 140 KJ=1,3
               BMSTAR(KJ)=BMSTAR(KJ) + BSTAR(KJ)
  140       CONTINUE
  100 CONTINUE
      DO 150 KH=1,3
         BMSTAR(KH)=BMSTAR(KH)/FLOAT(IB)
  150 CONTINUE
      DO 160 KI=1,3
         MSEBS=MSEBS+ BMSTAR(KI)*BMSTAR(KI)
  160 CONTINUE
      BLREG=MSEBS
      IF(ANS.EQ.1.AND.WI.EQ.1) WRITE(21,102) BLREG
  102 FORMAT(F8.4)
      RETURN
      END
C
      REAL FUNCTION XMEAN(X,NEK,WI)
      COMMON IB,ANS
      REAL X(1),Y(1000), V(10),BB(1000)
      INTEGER WI
      DO 10 I=1,NEK
         Y(I)=X(I)
   10 CONTINUE
      DO 15 I=1,IB
         DO 20 JI=1,NEK
            X(JI)=Y(JI)
   20    CONTINUE
         CALL BOOTS(X,NEK)
         CALL BSTATS(X,NEK,V)
         BB(I)= V(1)
   15 CONTINUE
      CALL BSTATS(BB,IB,V)
      XMEAN=V(1)
```

```
      IF(ANS.EQ.1.AND.WI.EQ.1) WRITE(21,102) XMEAN
  102 FORMAT(F8.4)
      RETURN
      END
C
C
      REAL FUNCTION VARIA(X,NEK,WI)
      COMMON IB,ANS
      REAL X(1), Y(1000),V(10),BB(1000)
      INTEGER WI
      DO 10 I=1,NEK
         Y(I)=X(I)
   10 CONTINUE
      DO 15 I=1,IB
         DO 20 JI=1,NEK
            X(JI)=Y(JI)
   20    CONTINUE
         CALL BOOTS(X,NEK)
         CALL BSTATS(X,NEK,V)
         BB(I)= V(2)
   15 CONTINUE
      CALL BSTATS(BB,IB,V)
      VARIA=V(1)
      IF(ANS.EQ.1.AND.WI.EQ.1) WRITE(21,102) VARIA
  102 FORMAT(F8.4)
      RETURN
      END
C
      REAL FUNCTION VARI2(X,NEK,WI)
      COMMON IB, ANS
      REAL X(1), Y(1000),V(10),BB(1000)
      INTEGER WI
      DO 10 I=1,NEK
         Y(I)=X(I)
   10 CONTINUE
      DO 15 I=1,IB
         DO 20 JI=1,NEK
            X(JI)=Y(JI)
   20    CONTINUE
         CALL BOOTS(X,NEK)
         CALL BSTATS(X,NEK,V)
         BB(I)= V(3)
   15 CONTINUE
      CALL BSTATS(BB,IB,V)
      VARI2=V(1)
      IF(ANS.EQ.1.AND.WI.EQ.1) WRITE(21,102) VARI2
  102 FORMAT(F8.4)
      RETURN
      END
C
      REAL FUNCTION VARI3(X,NEK,WI)
      COMMON IB, ANS
      REAL X(1), Y(1000),V(10),BB(1000),SMEAN,DNEK
      INTEGER WI
      DNEK=NEK
      SMEAN=0.0
      DO 10 I=1,NEK
         Y(I)=X(I)
         SMEAN=SMEAN+X(I)
   10 CONTINUE
      SMEAN=SMEAN/DNEK
      DO 15 I=1,IB
         DO 20 JI=1,DNEK
            X(JI)=Y(JI)
   20    CONTINUE
         CALL BOOTS(X,NEK)
         DO 30 JJ=1,NEK
         BB(I)= BB(I) + ((X(JJ)-SMEAN)**2)
   30    CONTINUE
         BB(I)=BB(I)/DNEK
   15 CONTINUE
      CALL BSTATS(BB,IB,V)
      VARI3=V(1)
      IF(ANS.EQ.1.AND.WI.EQ.1) WRITE(21,102) VARI3
  102 FORMAT(F8.4)
      RETURN
      END
C
      REAL FUNCTION COEVA(X,NEK,WI)
      COMMON IB,ANS
      REAL X(1), Y(1000),V(10),BB(1000)
      INTEGER WI
      DO 10 I=1,NEK
         Y(I)=X(I)
   10 CONTINUE
      DO 15 I=1,IB
         DO 20 JI=1,NEK
            X(JI)=Y(JI)
   20    CONTINUE
         CALL BOOTS(X,NEK)
         CALL BSTATS(X,NEK,V)
         BB(I)= V(4)
   15 CONTINUE
      CALL BSTATS(BB,IB,V)
      COEVA=V(1)
      IF(ANS.EQ.1.AND.WI.EQ.1) WRITE(21,102) COEVA
  102 FORMAT(F8.4)
      RETURN
      END
C
      REAL FUNCTION SECOR(X,NEK,WI)
      COMMON IB,ANS
      REAL X(1), Y(1000),V(10),BB(1000)
      INTEGER WI
      DO 10 I=1,NEK
         Y(I)=X(I)
   10 CONTINUE
      DO 15 I=1,IB
         DO 20 JI=1,NEK
            X(JI)=Y(JI)
   20    CONTINUE
```

```
          CALL BOOTS(X,NEK)
          CALL BSTATS(X,NEK,V)
          BB(I)= V(5)
   15 CONTINUE
      CALL BSTATS(BB,IB,V)
      SECOR=V(1)
      IF(ANS.EQ.1.AND.WI.EQ.1) WRITE(21,102) SECOR
  102 FORMAT(F8.4)
      RETURN
      END
C
      REAL FUNCTION MEDIA(X,NEK,WI)
      COMMON IB,ANS
      REAL X(1), Y(1000),V(10),BB(1000)
      INTEGER WI
      DO 10 I=1,NEK
         Y(I)=X(I)
   10 CONTINUE
      DO 15 I=1,IB
         DO 20 JI=1,NEK
            X(JI)=Y(JI)
   20    CONTINUE
         CALL BOOTS(X,NEK)
         CALL BSTATS(X,NEK,V)
         BB(I)= V(6)
   15 CONTINUE
      CALL BSTATS(BB,IB,V)
      MEDIA=V(1)
      IF(ANS.EQ.1.AND.WI.EQ.1) WRITE(21,102) MEDIA
  102 FORMAT(F8.4)
      RETURN
      END
C
      REAL FUNCTION TRIMM(X,NEK,WI)
      COMMON IB,ANS
      REAL X(1), Y(1000),V(10),BB(1000)
      INTEGER WI
      DO 10 I=1,NEK
         Y(I)=X(I)
   10 CONTINUE
      DO 15 I=1,IB
         DO 20 JI=1,NEK
            X(JI)=Y(JI)
   20    CONTINUE
         CALL BOOTS(X,NEK)
         CALL BSTATS(X,NEK,V)
         BB(I)= V(7)
   15 CONTINUE
      CALL BSTATS(BB,IB,V)
      TRIMM=V(1)
      IF(ANS.EQ.1.AND.WI.EQ.1) WRITE(21,102) TRIMM
  102 FORMAT(F8.4)
      RETURN
      END
C
C     BOOTSTRAP        SECTION
C
      SUBROUTINE BOOTS(X,NEK)
      COMMON IX4
      REAL X(1), XB(1000), XX(1000)
      CALL SRND(IX,XB,NEK,2,0)
      DO 10 I=1,NEK
         A=XB(I)
         B= A*NEK
         M=INT(B+1)
         IF(M.GT.NEK)M=NEK
         XX(I)=X(M)
   10 CONTINUE
      DO 20 I=1,NEK
         X(I)=XX(I)
   20 CONTINUE
      RETURN
      END
C
C     STATISTICS       SECTION
C
      SUBROUTINE BSTATS(X,NEK,V)
      COMMON IB
      REAL X(1), V(10), ZW(5000),ZT(5000),R,BMDIAN
      REAL*8 XMEAN,SUM2,SUM3,SUM4,SUUM4,DEV,DVAR,VSTD,DNB,SCOR
      REAL*8 XTRIM,BTRIM,VARIA2
      INTEGER BTAIL,ETAIL
C --- COMPUTE MEAN, STND DEVIATION, SKEWNESS, KURTOSIS, VARIANCE, CV
C---- MEDIAN, CORRELATION COEFF, AND TRIM(.05) MEAN.
      NB=NEK
      IF(NB.GT. 1) GO TO 10
      WRITE(6,100) NB
  100  FORMAT(2X,'SUBSAMPLE SEIZE IS TOO SMALL',F6.2)
       RETURN
   10 CONTINUE
      XMEAN=0.0
      DNB=NB
      DO 20 I=1,NB
        XMEAN=XMEAN+X(I)
   20 CONTINUE
      XMEAN=XMEAN/DNB
      V(1)=XMEAN
C     TO GENERATE HIGHER MOMENTS
      SUM2 = 0.0D0
      SUM3 = 0.0D0
      SUM4 = 0.0D0
      DO 30 I=1,NB
        DEV = X(I) - XMEAN
        SUM2 = SUM2 + DEV ** 2
        SUM3 = SUM3 + DEV ** 3
        SUM4 = SUM4 + DEV ** 4
   30 CONTINUE
C     BOOTSTRAP VARIANCE AND ITS STANDARD DEVIATION.
      DVAR = SUM2 / (DNB - 1.0D0)
      V(2)=DVAR
      VSTD=DSQRT(DVAR)
```

APPENDIX C

MSE_*^h OF SOME ESTIMATORS USING THE BOOTSTRAP METHOD

EST. MSE Of The Sample Mean Of An EXP(1)

B/n	10	20	25	40	50	70	100	140
5	0.1213	0.0544	0.0531	0.0309	0.0257	0.0216	0.0142	0.0118
8	0.1157	0.0570	0.0446	0.0299	0.0277	0.0164	0.0123	0.0103
10	0.1131	0.0551	0.0453	0.0288	0.0247	0.0170	0.0134	0.0097
15	0.1095	0.0543	0.0451	0.0277	0.0241	0.0164	0.0113	0.0099
20	0.1064	0.0528	0.0432	0.0262	0.0252	0.0163	0.0131	0.0096
25	0.1051	0.0525	0.0405	0.0270	0.0244	0.0153	0.0132	0.0097
40	0.1022	0.0508	0.0417	0.0277	0.0245	0.0162	0.0122	0.0087
60	0.1031	0.0511	0.0410	0.0258	0.0239	0.0159	0.0117	0.0091
100	0.1030	0.0512	0.0420	0.0252	0.0244	0.0155	0.0119	0.0090
140	0.1018	0.0511	0.0406	0.0256	0.0242	0.0156	0.0117	0.0092
500	0.1007	0.0471	0.0368	0.0217	0.0202	0.0119	0.0101	0.0041

EST. MSE Of The Sample Variance Of An EXP(1)

5	0.9130	0.5313	0.4114	0.1690	0.1703	0.1120	0.0745	0.1363
8	0.7783	0.4765	0.4023	0.1951	0.1538	0.1176	0.0847	0.0791
10	0.7776	0.5418	0.4485	0.1703	0.1461	0.1393	0.0680	0.0800
15	0.6732	0.5385	0.3457	0.1533	0.1433	0.1096	0.0650	0.0817
20	0.6408	0.4589	0.3447	0.1562	0.1373	0.1043	0.0662	0.0852
25	0.7115	0.4840	0.3452	0.1730	0.1311	0.0945	0.0656	0.0887
40	0.6822	0.4692	0.3392	0.1556	0.1349	0.1179	0.0635	0.0808
60	0.6959	0.4563	0.3265	0.1529	0.1341	0.1006	0.0658	0.0827
100	0.6857	0.4668	0.3434	0.1555	0.1285	0.1185	0.0643	0.0753
140	0.6789	0.4714	0.3259	0.1565	0.1280	0.1069	0.0592	0.0733
500	0.6649	0.4603	0.3035	0.1429	0.1098	0.0937	0.0394	0.0563

EST. MSE Of The Sample Coeff. of Variation Of An EXP(1)

5	0.0667	0.0391	0.0285	0.0238	0.0183	0.0144	0.0090	0.0080
8	0.0618	0.0352	0.0299	0.0249	0.0160	0.0156	0.0079	0.0080
10	0.0618	0.0340	0.0269	0.0218	0.0169	0.0126	0.0084	0.0080
15	0.0598	0.0336	0.0268	0.0221	0.0158	0.0127	0.0076	0.0079
20	0.0599	0.0313	0.0263	0.0218	0.0156	0.0133	0.0077	0.0068
25	0.0590	0.0323	0.0246	0.0223	0.0156	0.0137	0.0079	0.0074
40	0.0584	0.0309	0.0255	0.0208	0.0153	0.0120	0.0073	0.0071
60	0.0578	0.0313	0.0253	0.0214	0.0154	0.0127	0.0078	0.0070
100	0.0580	0.0304	0.0249	0.0213	0.0151	0.0122	0.0070	0.0073
140	0.0573	0.0308	0.0252	0.0215	0.0147	0.0123	0.0074	0.0074
500	0.0419	0.0297	0.0204	0.0187	0.0115	0.0100	0.0057	0.0039

Figure C.1 MSE_*^h of the Estimators for Exp(1).

B/n	5	10	15	20	25	30	50	60
5	0.4213	0.2045	0.2934	0.3217	0.1813	0.1527	0.0790	0.0565
8	0.4229	0.1951	0.2726	0.2332	0.1633	0.1383	0.0646	0.0449
10	0.3397	0.2134	0.2294	0.2195	0.1672	0.1376	0.0704	0.0417
15	0.3410	0.1904	0.2629	0.1974	0.1834	0.1415	0.0642	0.0442
20	0.3668	0.1975	0.2420	0.2365	0.1647	0.1467	0.0676	0.0430
25	0.3505	0.1859	0.2397	0.2229	0.1535	0.1067	0.0701	0.0437
30	0.3792	0.1851	0.2446	0.2307	0.1580	0.1196	0.0743	0.0449
35	0.3409	0.1927	0.2254	0.2228	0.1523	0.1234	0.0733	0.0438
40	0.3465	0.1896	0.2453	0.1988	0.1623	0.1215	0.0672	0.0426
45	0.3571	0.1852	0.2544	0.2191	0.1603	0.1290	0.0677	0.0420
50	0.3678	0.1888	0.2405	0.2318	0.1478	0.1191	0.0693	0.0439
100	0.3313	0.1785	0.2230	0.2191	0.1576	0.1229	0.0674	0.0409
500	0.3165	0.1582	0.1341	0.1217	0.1117	0.1095	0.0441	0.0287

EST. MSE Of The Sample Variance Of A N(0,1)

5	0.4158	0.2142	0.1416	0.1145	0.0987	0.0719	0.0413	0.0375
8	0.3841	0.2049	0.1363	0.1005	0.0970	0.0701	0.0490	0.0271
10	0.3650	0.1931	0.1346	0.1018	0.0930	0.0590	0.0424	0.0350
15	0.3687	0.1948	0.1332	0.1008	0.0853	0.0633	0.0444	0.0356
20	0.3541	0.1848	0.1298	0.0988	0.0835	0.0610	0.0420	0.0306
25	0.3712	0.1870	0.1225	0.0948	0.0848	0.0674	0.0398	0.0304
30	0.3570	0.1820	0.1250	0.0963	0.0847	0.0611	0.0416	0.0313
35	0.3632	0.1869	0.1266	0.0925	0.0850	0.0623	0.0399	0.0297
40	0.3474	0.1831	0.1252	0.0908	0.0818	0.0622	0.0414	0.0301
45	0.3595	0.1839	0.1223	0.0924	0.0809	0.0640	0.0408	0.0306
50	0.3625	0.1897	0.1211	0.0916	0.0827	0.0603	0.0408	0.0302
100	0.3644	0.1611	0.1132	0.0841	0.0806	0.0619	0.0412	0.0300
500	0.3175	0.1392	0.1008	0.0610	0.0715	0.0522	0.0391	0.0205

EST. MES Of The Sample Variance Of A L(0,1)

5	2.9553	2.3940	1.5890	1.0396	0.8608	0.7340	0.5076	0.4655
8	2.8503	2.0733	1.6019	0.9700	0.7033	0.6355	0.5318	0.3749
10	2.7371	2.0438	1.6862	0.9944	0.7115	0.7020	0.4938	0.4011
15	2.7377	1.9280	1.7109	0.9290	0.7775	0.6838	0.4844	0.3128
20	2.7954	1.8716	1.5557	0.9623	0.6811	0.6798	0.4974	0.3277
25	2.6397	1.8955	1.5850	0.9498	0.7466	0.6352	0.4633	0.3654
30	2.6941	1.8366	1.7492	0.8812	0.7106	0.6430	0.4849	0.3270
35	2.7119	1.8774	1.5792	0.8772	0.7000	0.6618	0.4890	0.3512
40	2.6518	1.8689	1.8452	0.8875	0.7028	0.6250	0.4785	0.3479
45	2.6200	1.8315	1.6082	0.9156	0.7119	0.5982	0.4987	0.3234
50	2.6419	1.8801	1.7016	0.8712	0.6749	0.6377	0.4652	0.3489
100	2.6334	1.8705	1.4931	0.8678	0.6827	0.6336	0.4763	0.3329
500	2.4163	1.6915	1.3852	0.7542	0.6173	0.5918	0.4258	0.3039

Figure C.2 MSE_*^h of S^2.

EST. MSE Of Sample Variance of a N(0,1)

B/n	5	10	15	20	25	30	50	60
5	0.4206	0.2099	0.1609	0.1025	0.0916	0.0680	0.0477	0.0379
8	0.3855	0.2032	0.1294	0.1084	0.0875	0.0702	0.0474	0.0316
10	0.3939	0.1986	0.1396	0.0964	0.0990	0.0667	0.0445	0.0292
15	0.3743	0.1942	0.1344	0.0961	0.0842	0.0658	0.0398	0.0325
20	0.3674	0.1854	0.1218	0.0971	0.0842	0.0665	0.0403	0.0319
25	0.3589	0.1898	0.1313	0.0968	0.0859	0.0619	0.0408	0.0312
30	0.3547	0.1851	0.1273	0.0949	0.0849	0.0615	0.0389	0.0317
35	0.3647	0.1861	0.1242	0.0949	0.0819	0.0622	0.0422	0.0310
40	0.3490	0.1851	0.1277	0.0928	0.0854	0.0631	0.0399	0.0314
45	0.3568	0.1871	0.1231	0.0915	0.0857	0.0632	0.0389	0.0298
50	0.3549	0.1862	0.1234	0.0940	0.0835	0.0650	0.0388	0.0311

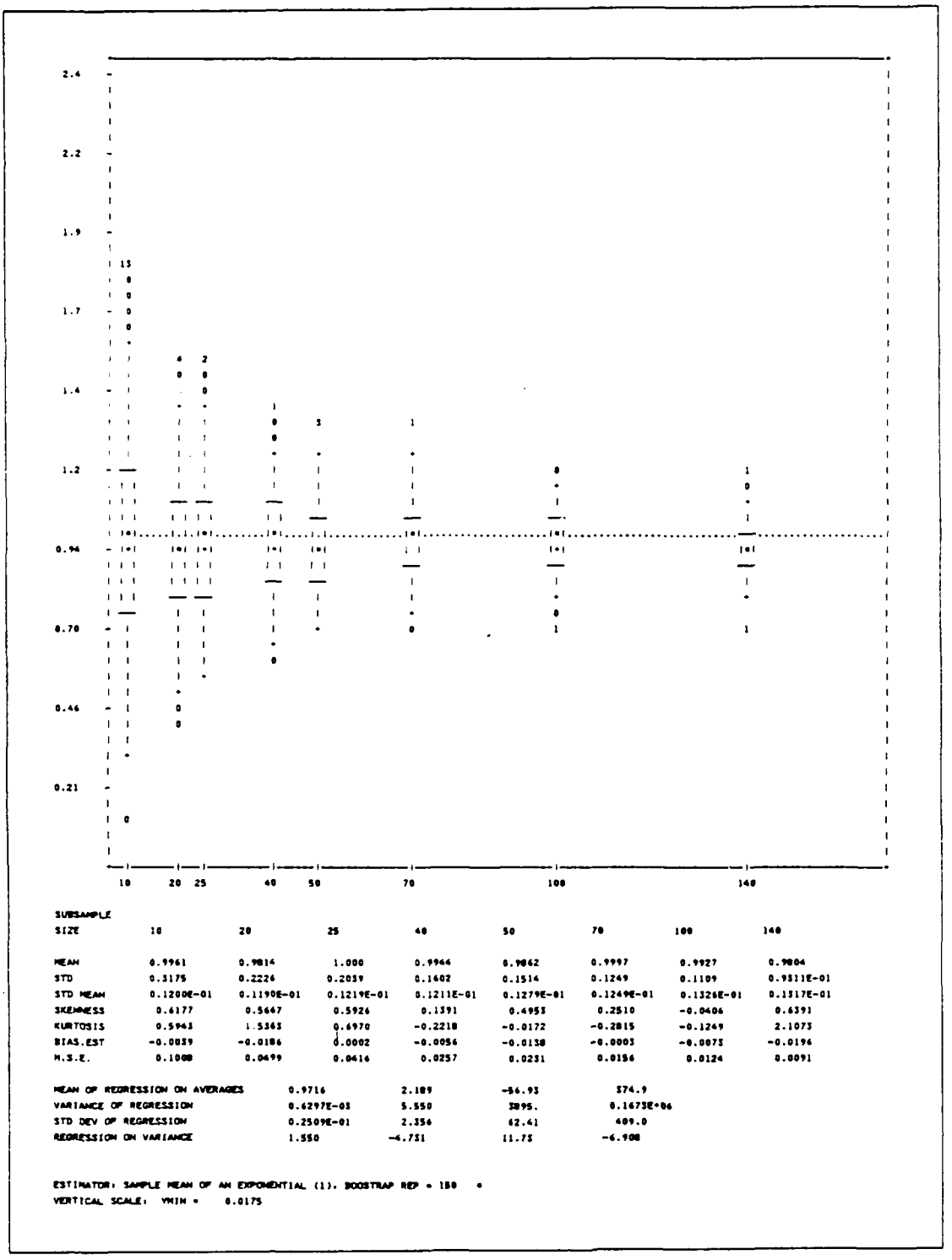
EST. MSE Of Sample Variance (2nd Estimator) of N(0,1)

5	0.5810	0.2467	0.1537	0.1091	0.0908	0.0879	0.0557	0.0384
8	0.5356	0.2540	0.1367	0.1164	0.0882	0.0844	0.0630	0.0368
10	0.5686	0.2461	0.1394	0.1026	0.0732	0.0790	0.0576	0.0408
15	0.5387	0.2304	0.1398	0.1067	0.0812	0.0685	0.0573	0.0369
20	0.5403	0.2285	0.1277	0.1043	0.0786	0.0727	0.0493	0.0383
25	0.5198	0.2204	0.1322	0.0989	0.0784	0.0754	0.0530	0.0340
30	0.5407	0.2270	0.1342	0.1023	0.0778	0.0742	0.0535	0.0330
35	0.5355	0.2249	0.1313	0.1005	0.0782	0.0740	0.0531	0.0347
40	0.5310	0.2225	0.1324	0.1034	0.0757	0.0744	0.0544	0.0356
45	0.5166	0.2261	0.1312	0.1036	0.0775	0.0713	0.0518	0.0362
50	0.5141	0.2242	0.1293	0.0994	0.0769	0.0712	0.0530	0.0360

EST. MSE Of Sample Variance (3rd Estimator) of a N(0,1)

5	0.3794	0.1714	0.1354	0.1222	0.0904	0.0673	0.0433	0.0410
8	0.3518	0.1706	0.1349	0.1173	0.0768	0.0612	0.0453	0.0363
10	0.3471	0.1729	0.1359	0.1132	0.0856	0.0622	0.0475	0.0403
15	0.3356	0.1542	0.1275	0.1055	0.0750	0.0578	0.0433	0.0364
20	0.3319	0.1568	0.1241	0.1119	0.0755	0.0595	0.0370	0.0345
25	0.3243	0.1615	0.1256	0.1089	0.0782	0.0563	0.0409	0.0332
30	0.3218	0.1573	0.1180	0.1095	0.0757	0.0552	0.0419	0.0322
35	0.3244	0.1576	0.1218	0.1034	0.0787	0.0553	0.0428	0.0320
40	0.3253	0.1522	0.1225	0.1076	0.0771	0.0553	0.0420	0.0366
45	0.3200	0.1573	0.1232	0.1056	0.0758	0.0565	0.0407	0.0351
50	0.3308	0.1565	0.1220	0.1064	0.0764	0.0552	0.0401	0.0347

Figure C.3 $MSE_*{}^h$ of ${}_1S^{*2}$, ${}_2S^{*2}$ and of ${}_3S^{*2}$.

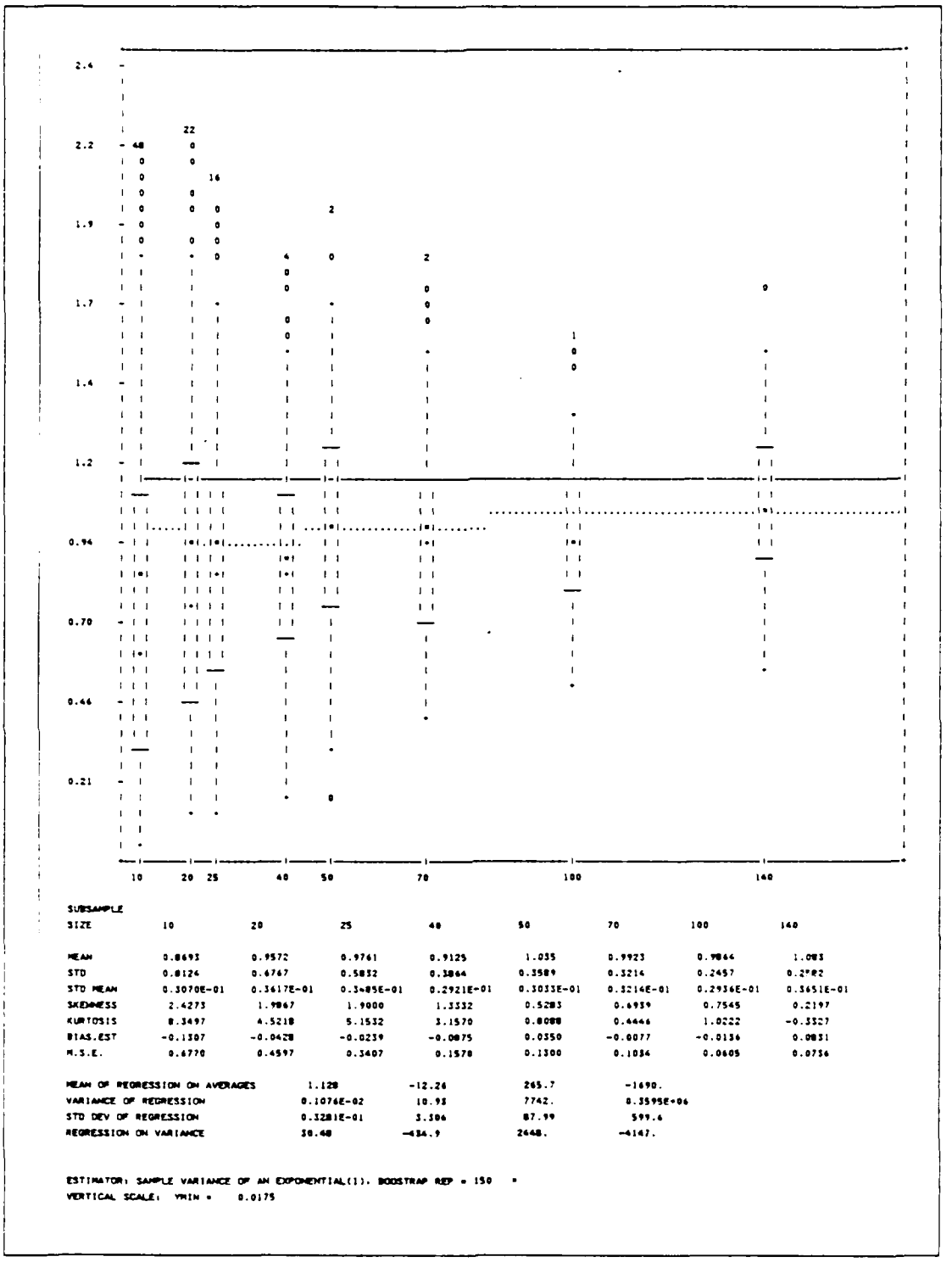


SUBSAMPLE SIZE	10	20	25	40	50	70	100	140
MEAN	0.9961	0.9814	1.000	0.9944	0.9862	0.9997	0.9927	0.9804
STD	0.3175	0.2226	0.2039	0.1602	0.1514	0.1249	0.1109	0.9311E-01
STD MEAN	0.1200E-01	0.1190E-01	0.1219E-01	0.1211E-01	0.1279E-01	0.1249E-01	0.1326E-01	0.1317E-01
SKEWNESS	0.6177	0.5667	0.5926	0.1391	0.4953	0.2510	-0.0406	0.6391
KURTOSIS	0.5943	1.5363	0.6970	-0.2218	-0.0172	-0.2815	-0.1249	2.1073
BIAS.EST	-0.0039	-0.0186	0.0002	-0.0056	-0.0138	-0.0003	-0.0073	-0.0196
M.S.E.	0.1008	0.0499	0.0416	0.0257	0.0231	0.0156	0.0124	0.0091

MEAN OF REGRESSION ON AVERAGES	0.9716	2.189	-56.93	574.9
VARIANCE OF REGRESSION	0.6297E-03	5.550	3895.	0.1675E+06
STD DEV OF REGRESSION	0.2509E-01	2.356	62.41	409.0
REGRESSION ON VARIANCE	1.550	-4.751	11.75	-6.908

ESTIMATOR: SAMPLE MEAN OF AN EXPONENTIAL (1). BOOSTRAP REP = 150 *
VERTICAL SCALE: YMIN = 0.0175

Figure C.4 Bootstrap Dist. of Sample Mean B = 150.



SUBSAMPLE SIZE	10	20	25	40	50	70	100	140
MEAN	0.8693	0.9572	0.9761	0.9125	1.035	0.9923	0.9844	1.003
STD	0.8124	0.6767	0.5832	0.3864	0.3589	0.3214	0.2457	0.2*02
STD MEAN	0.3070E-01	0.3617E-01	0.3+85E-01	0.2921E-01	0.3033E-01	0.3214E-01	0.2936E-01	0.3651E-01
SKEWNESS	2.4273	1.9867	1.9000	1.3332	0.5283	0.6939	0.7545	0.2197
KURTOSIS	8.3497	4.5218	5.1532	3.1570	0.8088	0.4446	1.0222	-0.3527
BIAS.EST	-0.1307	-0.0428	-0.0239	-0.0075	0.0350	-0.0077	-0.0136	0.0031
M.S.E.	0.6770	0.4597	0.3407	0.1578	0.1300	0.1034	0.0605	0.0736

MEAN OF REGRESSION ON AVERAGES	1.128	-12.26	265.7	-1690.
VARIANCE OF REGRESSION	0.1076E-02	10.93	7742.	0.3595E+06
STD DEV OF REGRESSION	0.3281E-01	3.306	87.99	599.6
REGRESSION ON VARIANCE	30.48	-434.9	2448.	-4147.

ESTIMATOR: SAMPLE VARIANCE OF AN EXPONENTIAL(1). BOOTSTRAP REP = 150 *
VERTICAL SCALE: YMIN = 0.0175

Figure C.5 Bootstrap Dist. of Sample Variance B = 150.

LIST OF REFERENCES

1. Efron, Bradley and Gong, Gail, *A Leisurely Look at the Bootstrap, the Jacknife, and Cross-Validation*, The American Statistician, February 1983, Vol. 37, No. 1, 36-48.

2. Miller, Rupert G., *The Jacknife-A Review*, Biometrika, 1974 , 61, 1-28.

3. Lewis, P.A.W., *Data Analysis and Simulation*, an unpublished work.

4. Efron, Bradley, *Bootstrap Method: Another Look at The Jacknife*, The Annals of Statistics, 1979, Vol.7, No.1, 1-26.

5. Efron, Bradley, *The Jacknife the Bootstrap, and Other Resampling Plans*, Society of Industrial and Applied Mathematics, 1982.

6. Efron, Bradley, *Censored Data and The Bootstrap*, Journal of the American Statistical Association, June 1981, Vol.76, No.3, 312-329.

7. Stanford University, Department of Statistics, Technical Report No.3, *Bootstrap Confidence Intervals*, by Robert Tibshirani, October 1984.

8. Lehman, E.L., *Theory Of Point Estimation* , Probability and Mathematical Statistics Series, Wiley, 1983.

9. Lewis, P.A.W., Orav, E.J., and Uribe, Luis, *Advanced Simulation and Statistics Package*, Wardworth and Brooks, 1986.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 23304-6145	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, California 93943-5000	2
3.	Commandant, USALMC ATTN:AMXMC-LS-S (MAJ McGram) FT. Lee, Virginia 23801-6040	2
4.	Prof. Donald R. Barr Naval Postgraduate School (Code 55Bn) Operation Research Department Monterey, California 93943-5000	2
5.	Prof. Toke Jayachandran Naval Postgraduate School (Code 53Jy) Department of Mathematics Monterey, California 93943-5000	2
6.	Commandant, USALMC ATTN:AMXMC-LS-S (CPT(P) Cortes-Colon) FT. Lee, Virginia 23801-6040	10

END

2-81

DTIC