

# Chapter 4

# Resampling Methods for Inference

February 26, 2001

Part of the **Iowa State University** NSF/ILI project

*Beyond Traditional Statistical Methods*

Copyright 1999, 2000, 2001 D. Cook, W. M. Duckworth, M. S. Kaiser, W. Q. Meeker and W. R. Stephenson.

Developed as part of NSF/ILI grant DUE9751644.

## Objectives

This chapter explains

- the use of simulation to study the behavior of sampling statistics,
- how simulation of sampling statistics motivates the Bootstrap methods, and
- how Bootstrap methods can be used to provide standard error estimates and confidence limits in cases where analytical solutions are difficult or impossible.

## Overview

Simulation is a powerful, modern tool that can be used to investigate the behavior of any estimation procedure where one is willing to completely specify the population being sampled. In these settings, one can obtain estimates for standard errors of parameter estimators even when the standard error formulas have not been determined analytically. Also, one can determine confidence intervals for unknown parameter values via simulation.

When one is unwilling or unable to completely specify the population under study, the “Bootstrap” allows one to obtain standard error estimates and/or confidence intervals. The Bootstrap is a general method (actually, a collection of methods) which uses the information contained in a single sample from the population of interest in conjunction with simulation results to provide information about the distribution of a sampling statistic. The Parametric Bootstrap requires one to partially specify the population under study by assuming a particular family of probability distributions although parameters specific to the chosen family are estimated. The Nonparametric Bootstrap doesn’t require any explicit assumptions about the population’s distribution but uses the single sample to provide an approximation to the population’s distribution.

## 4.1 Introduction

One of the primary activities of statisticians is estimation. When studying a population, one is often interested in one or more “characteristics” of the population; these characteristics are called *parameters*. The usual scenario is needing to estimate the value of a parameter based on the information contained in a single sample from the population of interest. This is the scenario addressed in this chapter.

The simplest and most studied case of estimating a population parameter is the case of estimating the *mean*,  $\mu$ . In some settings, the mean is not the only parameter of interest or, perhaps, not of interest at all. Some common parameters of interest include the percentiles of a population (including the 50th percentile—the median), standard deviation, skewness, kurtosis, correlation (with another population), and regression coefficients.

While the methods of this chapter are quite general and can be applied to almost any parameter estimation problem, one can benefit from seeing how these new methods compare to and fit in with the traditional methods introduced in elementary statistics courses.

## 4.2 A Familiar Setting: Estimating the Mean

Let  $\mu$  denote the (unknown) mean of the population under study, and let  $X_1, X_2, \dots, X_n$  denote a random sample from the population.

The problem of estimating the mean is typically handled by considering three different situations with varying degrees of information about the population available in each case. The three cases can be summarized as follows: “The data are independently sampled from. . .

- . . . a Normal population with (known) standard deviation  $\sigma$ .”
- . . . a Normal population with an unknown standard deviation.”
- . . . an “unknown” population with an unknown standard deviation.”

The phrase “unknown population with an unknown standard deviation” is a vague way of saying one doesn’t know anything about the population beyond the fact that it is believed to have a *finite* standard deviation. (Some populations do not have finite standard deviations.)

For each of these cases, the desired solution is not only an estimate but a confidence interval for the unknown mean.

### 4.2.1 A Normal Population With Known Standard Deviation

When estimating the mean of a Normal population with a known standard deviation  $\sigma$ , one relies on the fact that  $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$  can be shown to be a Normal random variable with mean  $\mu$  and standard error  $\text{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ . (The standard deviation of a sampling statistic is usually referred to as *standard error*. A *sampling statistic* is a statistic whose value varies with repeated sampling.) This fact leads to the usual confidence interval for the mean.

$$\begin{aligned} 1 - \alpha &= \Pr\left(\mu + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) \\ &= \Pr\left(z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) \\ &= \Pr\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Since the standard Normal distribution is symmetric about 0,  $-z_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}}$ . This allows the simplified notation  $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$  for the confidence interval  $[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$ . The notation  $z_{\frac{\alpha}{2}}$  denotes the  $(\frac{\alpha}{2} \times 100)$ th percentile of the standard Normal distribution with mean 0 and standard deviation 1; that is,  $\Pr(Z \leq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$  where  $Z$  is a standard Normal random variable.

The interpretation of the confidence interval is that  $(1 - \alpha) \times 100\%$  of *all samples* will result in confidence intervals which contain  $\mu$ . If a particular sample was randomly chosen from all possible samples, then one can be  $(1 - \alpha) \times 100\%$  confident that the resulting interval is one that contains  $\mu$ . Of course, there is also an  $\alpha \times 100\%$  chance that the particular sample leads to an interval which does *not* contain  $\mu$ , so unless we somehow discover the actual value of  $\mu$ , we will never know whether or not a particular interval contains the mean!

Notice that, in this setting, one can determine

- the mean of the estimator (the mean of  $\bar{X}$  is  $\mu$ ),
- the standard error of the estimator ( $\text{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ ), and
- the distribution of the estimator ( $\bar{X}$  is a Normal random variable).

### 4.2.2 A Normal Population With Unknown Standard Deviation

When estimating the mean of a Normal population with an unknown standard deviation, one must also estimate the standard deviation from the observed sample in order to obtain a confidence interval. Notice that the confidence interval really requires one to estimate (or know) the standard error of  $\bar{X}$  since the interval is of the form  $\bar{X} \pm z_{\frac{\alpha}{2}} \text{se}(\bar{X})$ ; however, the task is simplified by two facts:

- $\text{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- $\sigma$  can be estimated from the observed sample by

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

Although an estimate of  $\text{se}(\bar{X})$  is easily obtained from the observed sample, this solution introduces a new difficulty. Each possible sample leads to (potentially) different values of  $\bar{X}$  and  $s$ , so we have *two* random quantities associated with each sample. A resolution to this problem is reached by focusing on the quantity

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \quad (4.1)$$

and determining its distribution. In the previous setting where  $\sigma$  was known (and its value was used instead of the estimated value  $s$ ), this quantity had a standard Normal distribution. This was the reason for using the table values  $z_{\frac{\alpha}{2}}$  and  $z_{1-\frac{\alpha}{2}}$  in constructing the confidence interval for the mean.

One can show that (4.1) has a  $t$ -distribution with  $n - 1$  degrees of freedom. (This result depends on the fact that the sample was randomly selected from a Normal population.) This result leads to the usual confidence interval for this setting

$$\left[ \bar{X} - t_{1-\frac{\alpha}{2}}^{(n-1)} \frac{s}{\sqrt{n}}, \bar{X} - t_{\frac{\alpha}{2}}^{(n-1)} \frac{s}{\sqrt{n}} \right]$$

Since the  $t$ -distribution is symmetric about 0, we can express this interval in the simplified form  $\bar{X} \pm t_{\frac{\alpha}{2}}^{(n-1)} \frac{s}{\sqrt{n}}$  where  $t_{\frac{\alpha}{2}}^{(n-1)}$  denotes the  $(\frac{\alpha}{2} \times 100)$ th percentile of the  $t$ -distribution with  $n - 1$  degrees of freedom.

Notice that, in this setting, one can still determine the mean of the estimator (the mean of  $\bar{X}$  is  $\mu$ ) and the standard error of the estimator ( $\text{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ ); however, the confidence interval depends on

- having an estimator of  $\text{se}(\bar{X})$  (call the estimator  $\hat{\text{se}}(\bar{X})$ ) and
- knowing the distribution of

$$\frac{\bar{X} - \mu}{\hat{\text{se}}(\bar{X})} \quad (4.2)$$

In this setting,  $\hat{\text{se}}(\bar{X}) = \frac{s}{\sqrt{n}}$  and (4.2) is a  $t$  random variable with  $n - 1$  degrees of freedom.

### 4.2.3 An Unknown Population With an Unknown Standard Deviation

When estimating the mean of an “unknown” population with an unknown standard deviation, one relies on the Central Limit Theorem which states that (4.1) is approximately Normally distributed with a mean of 0 and an standard deviation of 1 for large  $n$  (as  $n$  increases, the approximation improves). The Central Limit Theorem is actually a collection of theorems stating various conditions under which (4.1) converges to a standard Normal distribution; however, the precise statements and convergence mechanisms of these theorems are beyond the scope of this chapter. The assumption of a finite standard deviation for the population under study is enough to allow the use of the Central Limit Theorem.

Even though the statistical theory behind this setting is more difficult than the first two settings, the construction of the confidence interval for the mean still depends on

- having an estimator of  $\text{se}(\bar{X})$  and
- knowing the (approximate) distribution of

$$\frac{\bar{X} - \mu}{\hat{\text{se}}(\bar{X})}$$

These facts lead to the *approximate* confidence interval  $[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}]$  or, simply,  $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ . The appropriateness of this interval depends on having a large sample.

Two things to note about this setting:

- If  $\sigma$  is known, one can use its value instead of the estimate  $s$ , and the Central Limit Theorem will still give us the approximate confidence interval as stated. However, for large  $n$ , one can show that there is little difference between knowing (and using)  $\sigma$  and using  $s$ .
- As  $n$  becomes large, a  $t$  random variable’s distribution is approximately a standard Normal distribution. So for large  $n$ , there is little difference between knowing one is sampling from a Normal population and not knowing the type of population from which one is sampling.

### 4.2.4 Summary

The three settings presented in this section should be familiar and are presented to serve as a review of elementary statistical concepts and as an introduction to some of the notation and terminology that will be used in this chapter. All three settings deal with the common problem of estimating the mean of a population,  $\mu$ , with the mean of a sample from the population,  $\bar{X}$ . Some important points about this estimation problem are:

- The formula for  $\text{se}(\bar{X})$  can be determined.
- $\text{se}(\bar{X})$  is easily estimated from the sample.
- The distribution of

$$\frac{\bar{X} - \mu}{\hat{\text{se}}(\bar{X})}$$

can be determined (at least approximately), and tables of percentiles for the distribution are available.

As mentioned in the introduction to this section, the mean is not the only parameter for which estimates and confidence intervals are often desired. Unfortunately, the mean is (almost) the only population parameter for which the above three points hold true without having to make restrictive (and sometimes unrealistic) assumptions about the population under study.

The methods of this chapter have the goal of providing a general, systematic approach to the estimation of population parameters other than the mean (although the methods presented can be used for the mean too). The general scenario will be as follows:

- A confidence interval for an (unknown) population parameter (denoted by  $\theta$ ) is desired.
- A sample from the population is available and denoted by  $(x_1, x_2, \dots, x_n)$  or  $\mathbf{x}$ .
- An estimate of  $\theta$  based on this sample is available. The estimate will be denoted as  $\hat{\theta}$  (or  $\hat{\theta}(\mathbf{x})$  to stress that its value is a function of the observed sample).

The details of how  $\hat{\theta}(\mathbf{x})$  is calculated from the observed sample will vary according to the available data, the parameter of interest, and the population under study. The methods of this chapter can be applied to any function of the data; however, how well the methods work depends (at least in part) on how “good” of an estimator  $\hat{\theta}$  is for  $\theta$ . One should at least be convinced that as the observed sample size  $n$  increases, the calculated value of  $\hat{\theta}(\mathbf{x})$  “approaches” the actual value of  $\theta$ . Also, a more computationally complex  $\hat{\theta}$  will greatly increase the computer time required to use the methods of this chapter.

### 4.3 Simulation of Sampling Statistics

The methods of this chapter are based, in part, on the concepts of simulation. As noted in Section 4.2, determining a confidence interval for a parameter requires some knowledge of the variability associated with the parameter’s estimator. In the case of the mean, we have  $\theta = \mu$  and  $\hat{\theta} = \bar{X}$ , and statistical theory provides answers to the questions of standard error and distribution of  $\hat{\theta}$  or at least the distribution of

$$\frac{\hat{\theta} - \theta}{\widehat{\text{se}}(\hat{\theta})} \quad (4.3)$$

However, when  $\theta$  is not the mean of the population and  $\hat{\theta}$  is not the mean of a sample, the needed standard error and distribution may be difficult or impossible to determine analytically. In situations like these, simulation can be used to gain an understanding of the estimation problem.

**Example 4.1** For the sake of illustrating the potential of simulation, this example considers the “problem” of estimating the 75th percentile of a standard Normal population; that is,  $\theta = z_{0.75}$ . Assume resources are available to obtain a sample of size  $n = 25$  from this population. For an estimator of  $\theta$ , the 75th percentile (the upper quartile) of the sample will be used. In a sample of size 25, the 75th percentile is  $X_{(19)}$  where the subscript (19) indicates the 19th *smallest* observed value (not necessarily  $X_{19}$ ), so  $\hat{\theta} = X_{(19)}$ .  $X_{(19)}$  is often called the 19th *order statistic*.

To see how  $\hat{\theta} = X_{(19)}$  behaves as an estimator of  $\theta = z_{0.75}$  for samples of size  $n = 25$ , one would like to see the results of *many* such samples from a standard Normal population (the assumed population of interest here). This is exactly what a simulation provides, and since the population of interest for this example is built-in to most statistics software packages, obtaining many such samples is relatively easy.

The following S+ commands simulate 1000 samples of size 25 from a standard Normal distribution. For each sample, the 75th percentile is calculated and stored. A histogram of the simulated 75th percentiles is displayed by the third command. The last two commands calculate the mean and standard deviation of the simulated percentiles.

```
y<-quantile(rnorm(25),0.75)
for(i in 1:999)
y<-c(y,quantile(rnorm(25),0.75))
hist(y)
mean(y)
sqrt(var(y))
```

Figure 4.1 displays one such histogram from a simulation. Of course, each time the code above is executed a different set of 1000 samples will be generated. Note the overall process used in the simulation:

1. Obtain a sample from the population, and

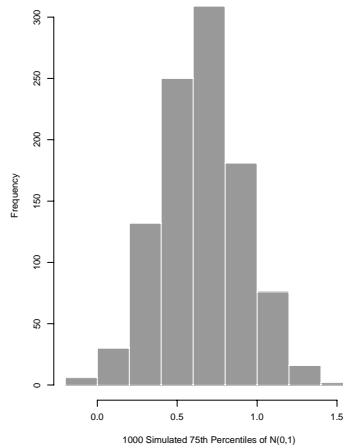


Figure 4.1: 1000 Simulated 75th Percentiles from a standard Normal population

2. calculate  $\hat{\theta}$ .
3. Repeat steps 1 and 2 obtaining  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_{1000}$ .
4. Make a histogram of the  $\hat{\theta}_i$ -values (and calculate the mean and standard deviation of these values).

At this point, the mean, standard error, and distribution of  $\hat{\theta}$  (a random quantity) are unknown; however, our simulation provides information about all three. The mean of the  $\hat{\theta}$  distribution ( $E(\hat{\theta})$ ) can be estimated by the mean of our 1000 simulated  $\hat{\theta}_i$ -values which is 0.6154768. Likewise, the standard error of  $\hat{\theta}$ ,  $se(\hat{\theta})$ , can be estimated by the standard deviation of our simulated percentiles which is 0.2616855. The histogram of the simulated  $\hat{\theta}_i$ -values provides an approximation to the distribution of  $\hat{\theta}$ . The histogram of 75th percentiles appears to be approximately Normal—one could also use a Normal Probability Plot to assess the Normality of our simulated percentiles.

One can use the simulated distribution of  $\hat{\theta}$  to obtain a confidence interval for the value of  $\theta$ . A reasonable confidence interval for  $\theta$  would be

$$[\hat{\theta}_{\frac{\alpha}{2}}, \hat{\theta}_{1-\frac{\alpha}{2}}] \quad (4.4)$$

where  $\hat{\theta}_{\frac{\alpha}{2}}$  is the  $(\frac{\alpha}{2} \times 100)$ th percentile of the simulated  $\hat{\theta}$  distribution. For a 95% confidence level ( $\alpha = 0.05$ ), the following S+ command (shown with its output) returns the required percentiles of our simulated 75th percentiles:

```
quantile(y,c(0.025,0.975))
      2.5%      97.5%
0.1246817 1.124073
```

In this special case, the value of  $\theta = z_{0.75}$  is known and can be found with the following command:

```
qnorm(.75)
0.6744898
```

The interval  $[0.1246817, 1.124073]$  does contain the actual value of  $\theta$ , 0.6744898. ■

The same methodology used in the example of this section (estimating the 75th percentile of the standard Normal population) can be used for parameters other than the 75th percentile and for populations other than the standard Normal. Also notice that even though  $\theta$  was known in this case, the distribution of  $\hat{\theta}$  (the 75th sample percentile) was unknown, and only through our simulation did we learn how our estimator is (approximately) distributed.

## 4.4 The Parametric Bootstrap Method

In Section 4.3, simulation of sampling statistics was presented as a general method for learning about the distribution of  $\hat{\theta}$ . Simulation is a very special case for several reasons:

- Enough is known about the population of interest that it can be completely specified; that is, the functional form of the distribution and all defining parameters are known.
- One can sample repeatedly from the population of interest (via computer).
- Usually (but not always) the actual value of  $\theta$  is known or can be determined analytically.

Even in these special circumstances where so much is known about the population being sampled, one may not be able to determine the distribution of  $\hat{\theta}$ . Simulation provides at least an approximate distribution for  $\hat{\theta}$ , and this distribution can be used to obtain approximate simulation-based confidence limits for the value of  $\theta$ .

The “Bootstrap” is a general method for obtaining information about the variability and distribution of  $\hat{\theta}$  and for obtaining confidence limits for  $\theta$  when the population of interest is *not* able to be completely specified. The Bootstrap method closely follows the ideas of simulation as presented in Section 4.3. When one knows (or is willing to assume that one knows) the functional form of the population under study but at least one defining parameter is unknown, the Bootstrap method is referred to as the *Parametric* Bootstrap method. If the distribution of the population is completely unknown, the Bootstrap method is referred to as the *Nonparametric* Bootstrap. This section covers the parametric scenario, and the next section covers the nonparametric scenario.

Recall that simulation involved repeatedly sampling from the known population to obtain many  $\hat{\theta}_i$ -values. The standard deviation of these  $\hat{\theta}_i$ -values provided an estimate of  $\text{se}(\hat{\theta})$ , the distribution of the  $\hat{\theta}_i$ -values provided an approximation to the distribution of  $\hat{\theta}$ , and the percentiles (or quantiles)  $\hat{\theta}_{\frac{\alpha}{2}}, \hat{\theta}_{1-\frac{\alpha}{2}}$  provided an approximate  $(1 - \alpha) \times 100\%$  confidence interval for  $\theta$ .

The Parametric Bootstrap is applied in situations where

- the functional form of the distribution is known,
- a single sample from the distribution is available, and
- the unknown parameters required to completely specify the distribution can be estimated from the single sample at hand.

**Example 4.2** Assume a single sample from a population believed to be Normally-distributed is available, and that the parameter of interest is  $\theta = x_{0.75}$  where  $x_{0.75}$  denotes the 75th percentile of this population. If the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of this population were known, this would be like Example 4.1, but both are unknown in this case. However, both can be estimated from the single sample at hand:  $\bar{x}$  provides an estimate of  $\mu$  and  $s$  provides an estimate of  $\sigma$ .

The Parametric Bootstrap method follows the simulation method with the exception that one repeatedly samples from a Normal population with a mean of  $\bar{x}$  and a standard deviation of  $s$ . Application of the Parametric Bootstrap method here requires that one have a single sample from the Normal population of interest (this sample is used to estimate the unknown defining parameters  $\mu$  and  $\sigma$  as well as the unknown parameter of interest  $\theta$ ) and a way to repeatedly sample from a Normal population with  $\mu = \bar{x}$  and  $\sigma = s$ .

Note the overall process used in this Parametric Bootstrap:

1. Obtain a single sample from the population and calculate  $\bar{x}$  and  $s$  as estimates of  $\mu$  and  $\sigma$ , respectively. These estimates *along with the assumption that the population is Normal* completely specify the population of interest (keep in mind that  $\mu$  and  $\sigma$  have been estimated though).
2. Obtain a sample (via computer) from a Normal population with a mean of  $\bar{x}$  and a standard deviation of  $s$ , and
3. calculate  $\hat{\theta}^*$ .

4. Repeat steps 2 and 3 obtaining  $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$ .
5. Make a histogram of the  $\hat{\theta}_i^*$ -values (and calculate the mean and standard deviation of these values).

The samples obtained in step 2 are called *bootstrap samples*, the  $\hat{\theta}_i^*$ -values obtained are called *bootstrap replicates*, and  $B$  stands for the number of bootstrap samples used. There is a subtle change in notation with respect to  $\hat{\theta}$  versus  $\hat{\theta}^*$ . In this setting, one can only obtain *one* sample from the population under study (the sample used in step 1). From this one sample, one estimate of  $\theta$  can be calculated—this estimate is denoted by  $\hat{\theta}$ . The  $B$  samples used in the rest of the process are *not* samples from the original population (Normal, mean  $\mu$ , standard deviation  $\sigma$ ) but from the “bootstrap” population (Normal, mean  $\bar{x}$ , standard deviation  $s$ ). For this reason, the resulting “ $\hat{\theta}_i^*$ -values” (step 3) from these samples are denoted by  $\hat{\theta}_i^*$ ; however, they are calculated in the same way that  $\hat{\theta}$  is calculated from the original sample.

The following S+ commands generate one sample from a Normal population with an unknown mean and unknown standard deviation and calculate the mean and standard deviation of this one sample. In this example, we *know* our single sample is from a Normal population because S+ has generated the sample from a Normal distribution, so this is a perfect candidate for the Parametric Bootstrap method. In practice, one may not *know* the original sample is from a Normal population, but one may be reasonably sure (or willing to assume) that the sample is from a Normal population. The Parametric Bootstrap requires that the form of the sampled population be specified, but it need not be Normal—Normal is being used for the sake of this example—one can specify any one of the many distributions built-in to S+.

```
mu<-runif(1,-10,10)
stdev<-runif(1,0.5,10.5)
x1<-rnorm(25,mu,stdev)
xbar<-mean(x1)
s<-sqrt(var(x1))
```

The following commands parallel the commands used in Section 4.3; however, the repeated sampling is from a Normal population with a mean of  $\bar{x}$  (rather than  $\mu$ ) and a standard deviation of  $s$  (rather than  $\sigma$ ) because  $\mu$  and  $\sigma$  are unknown in this scenario.

```
y<-quantile(rnorm(25,xbar,s),0.75)
for(i in 1:999)
y<-c(y,quantile(rnorm(25,xbar,s),0.75))
hist(y)
mean(y)
sqrt(var(y))
```

Figure 4.2 displays a histogram of  $\hat{\theta}_i^*$ -values generated by the commands above. Of course, each time the commands are executed a different  $B = 1000$  bootstrap samples will be generated.

The  $B$  bootstrap replicates  $\hat{\theta}_i^*$  provide the following information about the distribution of  $\hat{\theta}$ :

- The mean of the  $\hat{\theta}$  distribution is estimated by

$$\hat{\theta}^* = \frac{1}{B} \sum_{i=1}^B \hat{\theta}_i^* \quad (4.5)$$

- The standard error of the  $\hat{\theta}$  distribution is estimated by

$$\hat{s}e_B(\hat{\theta}) = \sqrt{\frac{\sum_{i=1}^B (\hat{\theta}_i^* - \hat{\theta}^*)^2}{B-1}} \quad (4.6)$$

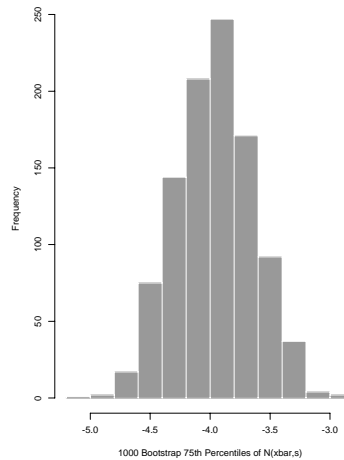


Figure 4.2: 1000 Bootstrap 75th Percentiles from a  $N(\bar{x},s)$  population

- The distribution of  $\hat{\theta}$  is approximated by the histogram of the  $\hat{\theta}_i^*$ -values, and the percentiles  $\hat{\theta}_{\frac{\alpha}{2}}^*$ ,  $\hat{\theta}_{1-\frac{\alpha}{2}}^*$  provide an approximate confidence interval for  $\theta$ .

The interval

$$[\hat{\theta}_{\frac{\alpha}{2}}^*, \hat{\theta}_{1-\frac{\alpha}{2}}^*] \quad (4.7)$$

is referred to as a *parametric bootstrap confidence interval for  $\theta$* . The following S+ command (shown with its output) provides the percentiles for a 95% confidence interval:

```
quantile(y,c(0.025,0.975))
 2.5%      97.5%
-4.546366 -3.297336
```

In this special case (the original sample was generated via computer), the actual value of  $\theta$  can be determined with the following S+ command (shown with its output):

```
qnorm(0.75,mu,stdev)
-3.822578
```

The bootstrap interval  $[-4.546366, -3.297336]$  contains  $\theta = -3.822578$ . ■

The Parametric Bootstrap method illustrated by this example can be applied to any situation where

- a single sample from a population is available,
- one is willing to specify the functional form of the distribution of the population (with the exception of certain defining parameters for the distribution which can be estimated from the original sample), and
- one has the ability to repeatedly sample from the specified population distribution (using estimated values of defining parameters).

While the previous example involved estimating the 75th percentile of a population known to be Normal, the method can be applied to obtain confidence intervals for *any* parameter  $\theta$  based on a sample from *any* population as long as one is willing to specify the functional form of the population. The situation where one is unable (or unwilling) to specify the functional form of the population distribution is the topic of Section 4.5.

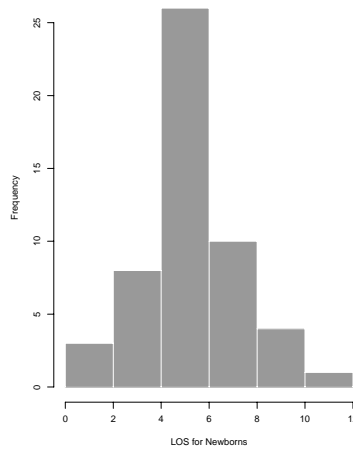


Figure 4.3: Length of Stay for 52 Newborns

**Example 4.3** The number of days in the hospital after birth was recorded for a random sample of 52 newborns (during one year at a particular hospital). If a confidence interval for the *mean* length of stay (LOS) for newborns was desired, then the Central Limit Theorem would allow use of the usual confidence interval based on percentiles of the standard Normal distribution since  $n \geq 30$ . However, assume that the parameter of interest is the *median* LOS for newborns. Figure 4.3 is a histogram of the 52 LOS observations.

From inspection of the histogram of LOS for this sample of 52 newborns, an assumption of Normality for the population of all newborn LOS values seems reasonable. Applying the Parametric Bootstrap method (see Example 4.2), one obtains the histogram of bootstrap replicates shown in Figure 4.4.

The histogram of bootstrap replicates of the median indicates that the distribution of the median of samples of size 52 from a Normal population is approximately Normal. The percentiles of this “bootstrap distribution” are given by

```
quantile(y,c(0.025,0.975))
 2.5%    97.5%
5.128986 6.502569
```

and constitute a 95% bootstrap confidence interval for the median LOS of newborns.

If one consults the records of all newborns for this specific hospital during this specific year, one finds that the median was 6 which is contained in the above interval. ■

## 4.5 The Nonparametric Bootstrap Method

Section 4.4 presented the Parametric Bootstrap method for learning about the distribution of a sampling statistic  $\hat{\theta}$ . The Parametric Bootstrap requires that one specify only the functional form of the distribution being sampled whereas the simulations presented in Section 4.3 require that one completely specify the distribution being sampled (functional form as well as the values of defining parameters). The Nonparametric Bootstrap method, often called simply “The Bootstrap”, goes one step beyond the Parametric Bootstrap method by not requiring that one explicitly specify anything about the population being sampled. All the Nonparametric Bootstrap requires is a single sample from the population being studied in order to provide standard error estimates and confidence intervals.

The overall process used in the Nonparametric Bootstrap is:

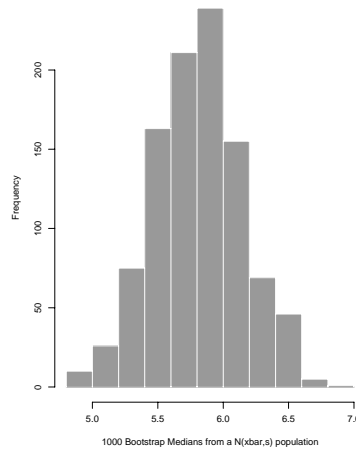


Figure 4.4: 1000 Bootstrap Medians from a  $N(\bar{x}, s)$  population

1. Obtain a single sample from the population under study and calculate  $\hat{\theta}$  (the estimate of  $\theta$  based on the sample).
2. Generate a bootstrap sample (from your original sample) by sampling *with replacement* from your original sample.
3. Calculate  $\hat{\theta}^*$ .
4. Repeat steps 2 and 3 obtaining  $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$ .
5. Make a histogram of the  $\hat{\theta}_i^*$ -values (and calculate the mean and standard deviation of these values).

The difference between the Nonparametric Bootstrap and the Parametric Bootstrap is the “population” sampled from in obtaining the bootstrap samples (step 2). In the parametric setting, the bootstrap samples are drawn from a specific distributional family (e.g. Normal, Exponential, etc.). In the nonparametric setting, the bootstrap samples are drawn *with replacement* from the *original sample*. This is equivalent to generating samples from the discrete empirical probability distribution given by the values  $x_1, x_2, \dots, x_n$ . This probability distribution assigns a probability of  $\frac{1}{n}$  to each observation. If a particular value appears exactly twice in the sample, say  $x_1 = x_2$ , then that common value has a probability of  $\frac{2}{n}$ . Essentially, one treats the relative frequencies derived from the original sample as the population’s relative frequencies and uses these relative frequencies to generate the bootstrap samples.

**Example 4.4** In Example 4.3, a sample of 52 newborns’ length of stays (LOS) was used to provide a parametric bootstrap confidence interval for the median LOS of (all) newborns. The Parametric Bootstrap was used because the histogram of the sample values (Figure 4.3) suggested a mound-shaped distribution. If one was unwilling to assume the sample came from a Normal population (or if Figure 4.3 did not have a “familiar” shape), then one could apply the Nonparametric Bootstrap method to obtain a confidence interval for median LOS.

The Nonparametric Bootstrap method is implemented in S+ by the function named `bootstrap`. The following commands use this function to obtain 1000 bootstrap replicates of the median from our original sample of 52 newborns. A histogram of the bootstrap medians is displayed and a 95% confidence interval for the population median is calculated. The original sample is named `x` in these commands.

```
y<-bootstrap(x,1000,median)
thetastars<-y[[1]]
hist(thetastars)
```

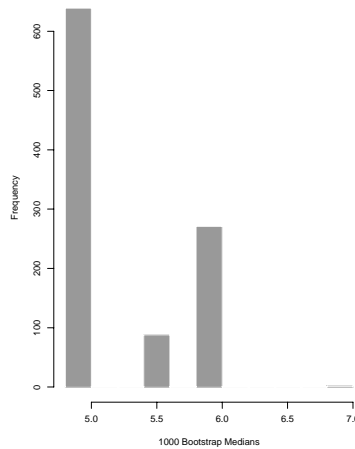


Figure 4.5: 1000 Bootstrap Medians

```
quantile(thetastars,c(0.025,0.975))
 2.5% 97.5%
 5     6
```

Figure 4.5 displays a histogram of the 1000 bootstrap medians. This histogram should be compared to that of Figure 4.4 which contains the 1000 bootstrap medians obtained from the Parametric Bootstrap method. The discrete nature of the “population” (the original sample) from which the bootstrap medians are calculated in the nonparametric approach is reflected in the histogram of those replicates. The median LOS for all newborns born in the particular year under study is 6 which is “on the edge” of our nonparametric bootstrap confidence interval. ■

## 4.6 Better Bootstrap Intervals

The bootstrap confidence intervals presented in Sections 4.4 and 4.5 are of the form  $[\hat{\theta}_{\frac{\alpha}{2}}^*, \hat{\theta}_{1-\frac{\alpha}{2}}^*]$  where  $\hat{\theta}_{\frac{\alpha}{2}}^*$  is the  $(\frac{\alpha}{2} \times 100)$ th percentile of the bootstrap replicates. This method of obtaining a confidence interval for  $\theta$  is called the *percentile method*. The percentile method of obtaining bootstrap confidence intervals is *intuitively* how the bootstrap replicates are used to obtain confidence limits; however, in practice, the actual coverage rates of percentile method intervals do not always match well with their target coverage rates. That is, 95% percentile method confidence intervals do not always contain  $\theta$  95% of the time. There are at least four proposed alternatives to the percentile method:

- the bootstrap- $t$  method,
- the variance-stabilized bootstrap- $t$  method,
- the  $BC_a$  method, and
- the ABC method.

With the exception of the bootstrap- $t$  method, the ideas behind these alternatives are beyond the scope of this chapter; however, all four methods are implemented in S+ and can easily be used to obtain better confidence intervals for  $\theta$ .

$BC_a$  stands for *bias-corrected and accelerated*. The  $BC_a$  method improves upon the percentile method both theoretically and in practice. The  $BC_a$  method does require more bootstrap replications to obtain

its confidence limits than the percentile method; however, the intervals are better with respect to coverage rates. The  $BC_a$  method is implemented in S+ with the function `bcanon`. ABC stands for *approximate bootstrap confidence*. The ABC method uses some analytical results to reduce the number of bootstrap samples needed in the  $BC_a$  method. The ABC method is implemented in S+ with the function `abcnon`.

The bootstrap- $t$  and the variance-stabilized bootstrap- $t$  methods both work off the common idea of how the standard  $t$ -intervals “correct” the standard  $z$ -intervals when the sample size is small (see Section 4.2). The standard  $z$ -interval is

$$[\hat{\theta} - z_{1-\frac{\alpha}{2}} \hat{s}\hat{e}(\hat{\theta}), \hat{\theta} - z_{\frac{\alpha}{2}} \hat{s}\hat{e}(\hat{\theta})]$$

where  $\hat{\theta} = \bar{X}$  and  $\hat{s}\hat{e}(\hat{\theta}) = \frac{s}{\sqrt{n}}$ . However, when the sample size is small, the  $z$ -percentiles used are not appropriate. The “correction” is to use the corresponding  $t$ -percentiles—this requires that the sampled population be Normally-distributed.

$$[\hat{\theta} - t_{1-\frac{\alpha}{2}}^{(n-1)} \hat{s}\hat{e}(\hat{\theta}), \hat{\theta} - t_{\frac{\alpha}{2}}^{(n-1)} \hat{s}\hat{e}(\hat{\theta})]$$

The bootstrap- $t$  method takes this idea one step further and frees one from the requirement of a Normal population.

Essentially, the idea of the bootstrap- $t$  method is to treat the values  $t_{1-\frac{\alpha}{2}}$  and  $t_{\frac{\alpha}{2}}$  as unknown parameters (percentiles are parameters) of the distribution of the values

$$t = \frac{\hat{\theta} - \theta}{\hat{s}\hat{e}(\hat{\theta})}$$

and use the bootstrap method to obtain  $B$  replicates of these values. The percentiles of the bootstrap replicates are then used as estimates of the unknown percentiles. Note: The letter  $t$  is being used to denote quantities that need not be related the usual  $t$ -distribution (unless  $\theta = \mu$  and  $\hat{\theta} = \bar{X}$ ). This is done in an effort to maintain the intuitive connection to the  $t$ -distribution while not introducing more complex notation.

In general,  $\hat{s}\hat{e}(\hat{\theta})$  is not known and is usually estimated via a second level of “bootstrapping”. The bootstrap- $t$  method involves the following process:

1. Obtain a single sample from the population under study and calculate  $\hat{\theta}$  (the estimate of  $\theta$  based on the sample).
2. Generate a bootstrap sample (from your original sample) by sampling with replacement from your original sample and calculate  $\hat{\theta}^*$ .
3. (a) Generate a bootstrap subsample (from your bootstrap sample) by sampling with replacement from your bootstrap sample and calculate  $\hat{\theta}^{**}$   
 (b) Repeat step 3(a) to obtain  $\hat{\theta}_1^{**}, \hat{\theta}_2^{**}, \dots, \hat{\theta}_{B_2}^{**}$  and use the standard deviation of the  $\hat{\theta}_i^{**}$ -values,  $\hat{s}\hat{e}_{B^*}(\hat{\theta})$ , as an estimate of  $\hat{s}\hat{e}(\hat{\theta})$ .  
 (c) Calculate

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{s}\hat{e}_{B^*}(\hat{\theta})}$$

4. Repeat steps 2 and 3 obtaining  $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$  and  $t_1^*, t_2^*, \dots, t_B^*$ .
5. Make a histogram of the  $t_i^*$ -values and use the percentiles  $t_{\frac{\alpha}{2}}^*, t_{1-\frac{\alpha}{2}}^*$  to obtain the confidence interval for  $\theta$ :

$$[\hat{\theta} - t_{1-\frac{\alpha}{2}}^* \hat{s}\hat{e}_B(\hat{\theta}), \hat{\theta} - t_{\frac{\alpha}{2}}^* \hat{s}\hat{e}_B(\hat{\theta})]$$

where  $\hat{s}\hat{e}_B(\hat{\theta})$  is, as usual, the standard deviation of the  $\hat{\theta}_i^*$ -values.

The bootstrap- $t$  method is implemented in S+ with the function `boott`. The variance-stabilized bootstrap- $t$  method corrects for a deficiency in the bootstrap- $t$  method and is implemented in S+ by using `boott` with the option `VS=T`.

**Example 4.5** A sample of 17 Ames' restaurants' health code ratings is used to provide a confidence interval for  $\theta$ , the Interquartile Range (IQR) of all Ames' restaurant ratings. The IQR is the 75th percentile minus the 25th percentile.

The following S+ commands (shown with their output) provide  $BC_a$  and bootstrap- $t$  intervals. The sample of 17 restaurants is `x` in the following commands.

```
iqr<-function(x){quantile(x,0.75)-quantile(x,0.25)}
y<-bcanon(x,1000,iqr)
y[[1]]
      alpha bca point
[1,] 0.025      4
[2,] 0.050      6
[3,] 0.100      6
[4,] 0.160      7
[5,] 0.840     11
[6,] 0.900     12
[7,] 0.950     13
[8,] 0.975     13

y<-boott(x,iqr)
y[[1]]
      0.001    0.01    0.025    0.05    0.1    0.5    0.9    0.95
[1,] 4.663441 5.559459 5.828737 6.75071 7.392436 9.541092 12.36138 13.01542
      0.975    0.99    0.999
[1,] 13.30755 14.6261 16.88917
```

The  $BC_a$  95% confidence interval for  $\theta$  is  $[4, 13]$ , and the bootstrap- $t$  95% confidence interval for  $\theta$  is  $[5.828737, 13.30755]$ . The actual IQR for all Ames' restaurant ratings is  $\theta = 9$ . ■

The bootstrap- $t$  method is generally good for estimating location parameters (like the median) especially in the Parametric Bootstrap setting, but is not as reliable at estimating other parameters especially for small samples. At present, the  $BC_a$  intervals seem to be the best "general purpose" bootstrap-based intervals especially in the Nonparametric Bootstrap setting. Some general guidelines on how many bootstrap replicates are needed for each method are:

$B$	method
25–200	$\hat{s}e_B(\hat{\theta})$ only
1000	Percentile interval
$1000 \times 25$	Bootstrap- $t$ interval
$(100 \times 25) + 1000$	Variance-stabilized bootstrap- $t$ interval
1000–2000	$BC_a$ interval
50–100	ABC interval

Of course, the larger  $B$  one uses the more accurate the methods are in general. If you have the time and computing power,  $B = 10000$  seems to be a popular number of replicates to use for most applications.

## 4.7 Bootstrap in Regression

A rough draft at best!

In regression, one usually makes distributional assumptions about the errors terms (e.g. Normal, mean zero, equal standard deviations). These assumptions are not always met by a particular data set, and one usually discovers this by looking at residual plots. There are at least two ways to apply the bootstrap concept to the regression setting.

Consider simple linear regression:  $y = \beta_0 + \beta_1 x + \epsilon$ . If one believes that  $\beta_0 + \beta_1 x$  is the appropriate functional form for the mean of  $y$ ; however, the  $\epsilon$  terms appear to violate their distributional assumptions, then one might consider bootstrapping the residuals to obtain more accurate information about their distribution and use the usual least squares estimates of  $\beta_0$  and  $\beta_1$  to filter this information back into the model (i.e. to get confidence intervals for the  $\beta_i$ 's). However, if one is unsure of the form of the model, then one might choose to bootstrap the original data  $(x_i, y_i)$  to obtain confidence intervals for the  $\beta_i$ 's directly.

**Example 4.6** This example involves the following data on the speed of a car (the first column) and the distance required for the car to stop (the second column). The original pairs will be bootstrapped rather than the residuals from the usual least-squares regression fit. This example also focuses on  $\beta_1$  only, but I'm sure there are more sophisticated ways to bootstrap  $\beta_0$  and  $\beta_1$  simultaneously using S+. Below are the contents of the file `reg.txt`.

```
20 15
30 34
40 73
50 110
60 152
```

The following S+ commands (shown with their output) provide the bootstrap replicates of  $\hat{\beta}_1$ .

```
reg<-matrix(scan(file="reg.txt"),ncol=2,byrow=T,dimnames=
list(1:5,c("Speed","StopDist")))
reg
  speed stopdist
1    20      15
2    30      34
3    40      73
4    50     110
5    60     152

cor(reg[,1],reg[,2])
0.9930207

beta1<-function(x,reg){
cor(reg[x,1],reg[x,2])*sqrt(var(reg[x,2]))/sqrt(var(reg[x,1]))}
n<-5
beta1(1:n,reg)
3.5
y<-bootstrap(1:n,1000,beta1,reg)
ysort<-sort(y[[1]])
quantile(ysort,c(0.025,0.975))
 2.5%      97.5%
 2.775   3.985714
motif()
hist(ysort)
```

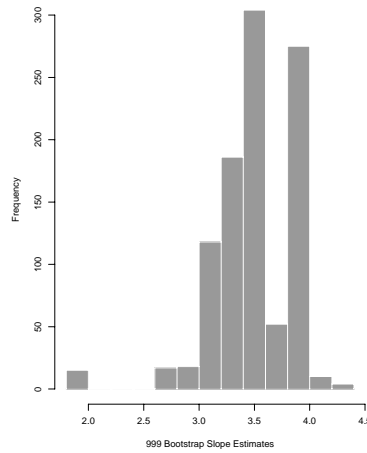


Figure 4.6: 999 Bootstrap Slope Estimates

Figure 4.6 displays a histogram of the bootstrap replicates of the slope estimate. One of the bootstrap samples consisted of 5 copies of the same point from the data set, so no slope estimate could be obtained from this bootstrap sample (that is why there are only 999 replicates plotted). The usual 95% confidence interval for  $\beta_1$  here is  $[2.7362, 4.2638]$  while our bootstrap-based 95% confidence interval is  $[2.775, 3.985714]$ . ■

## References

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM.

Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman & Hall.

## Exercises

**4.1** The methodology of Section 4.3 provides a mechanism for obtaining a confidence interval for an unknown population parameter  $\theta$  through simulation of the sampling distribution of  $\hat{\theta}$ . Section 4.2 reviews the analytically derived confidence intervals for the case of  $\theta = \mu$  and  $\hat{\theta} = \bar{X}$ . This exercise provides a comparison of the two approaches.

- (a). Use S+ to obtain a single sample of size 12 from a Normal population with an unknown mean and a standard deviation of 2. Based on this one sample, calculate a 95% confidence interval for the mean.

The following S+ commands generate such a sample, provide summary statistics, and give the value of  $z_{0.975}$  needed to calculate the confidence interval. Refrain from typing just `mu` in S+ as this will show you the “unknown” value of  $\mu$  prematurely.

```
mu<-runif(1,-10,10)
x<-rnorm(12,mu,2)
mean(x)
qnorm(0.975)
```

Note that  $\sigma$  is *known* in this problem and should be used in the confidence interval calculation.

- (b). If the process above was repeated many times, approximately 95% of the resulting intervals would contain the unknown value of  $\mu$ . The following S+ commands obtain 100 samples like the samples from part (a) and count how many of the resulting 95% confidence intervals actually contain the unknown value of  $\mu$ .

```
y<-0
for(j in 1:100){
  x<-rnorm(12,mu,2)
  lcl<-mean(x)-qnorm(0.975)*2/sqrt(12)
  ucl<-mean(x)+qnorm(0.975)*2/sqrt(12)
  if(lcl<mu && mu<ucl) y<-y+1}
print(y)
```

Note the number of confidence intervals that contain  $\mu$ . How does this result compare to the expected “coverage rate” of 95%?

By adding the line

```
for(i in 1:10){
```

at the beginning of the S+ commands and changing the last line to

```
print(y)}
```

one can repeat the above exercise 10 times to see how the coverage rates themselves vary.

- (c). The interval found in part (a), is based on the important fact that for estimating  $\mu$  with  $\bar{X}$  both  $\text{se}(\bar{X})$  and the distribution of  $\bar{X}$  are known. However, the simulation-based interval of Section 4.3 is not based on this type of theoretical knowledge of  $\hat{\theta}$ . The real test of the simulation method is not whether a single simulation-based confidence interval contains  $\mu$  but whether or not 95% of all such intervals contain  $\mu$ . The following S+ commands will generate 100 simulation-based confidence intervals for  $\mu$  (each based on 100 samples) and print how many of the 100 intervals actually contain (the still unknown value of)  $\mu$ .

```
y<-0
for(j in 1:100){
x<-mean(rnorm(12,mu,2))
for(i in 1:99){
x<-c(x,mean(rnorm(12,mu,2)))}
if(quantile(x,0.025)<mu && mu<quantile(x,0.975)) y<-y+1 }
print(y)
```

Note the number of simulation-based intervals that contained the unknown value of  $\mu$ . Does this result indicate that the simulation-based intervals are reasonable in this setting? Explain your response briefly.

- (d). Enter mu at the S+ prompt to reveal the actual value of  $\mu$  you have been trying to estimate and record that value.

**4.2** Repeat Exercise 4.1 with an unknown standard deviation. The following S+ commands provide the required results.

- (a). Calculate a single 95% confidence interval for the unknown mean.

```
mu<-runif(1,-10,10)
stdev<-runif(1,0.5,10.5)
x<-rnorm(12,mu,stdev)
mean(x)
sqrt(var(x))
qt(0.975,11)
```

- (b). Count the number of such intervals (out of 100) containing the unknown mean (repeated 10 times to see how the coverage rates vary).

```
for(i in 1:10){
y<-0
for(j in 1:100){
x<-rnorm(12,mu,stdev)
lcl<-mean(x)-qt(0.975,11)*sqrt(var(x))/sqrt(12)
ucl<-mean(x)+qt(0.975,11)*sqrt(var(x))/sqrt(12)
if(lcl<mu && mu<ucl) y<-y+1}
print(y)}
```

How do the coverage rates compare to the expected 95%?

- (c). Count the number of simulation-based intervals (out of 100) containing the unknown mean.

```

y<-0
for(j in 1:100){
x<-mean(rnorm(12,mu,stdev))
for(i in 1:99){
x<-c(x,mean(rnorm(12,mu,stdev)))}
if(quantile(x,0.025)<mu && mu<quantile(x,0.975)) y<-y+1 }
print(y)

```

How does the simulation-based method compare to the targeted 95% coverage rate?

**4.3** Repeat Exercise 4.1 with an “unknown” distribution. The population from which the samples are drawn for this exercise is the *Exponential*; however, familiarity with this distribution is not required. The sample size used is  $n = 30$  since the Central Limit Theorem is being relied upon to insure the validity of the confidence limits. The following S+ commands provide the required results.

- (a). Calculate a single 95% confidence interval for the unknown mean.

```

mu<-runif(1,0.25,2.25)
rate<-1/mu
x<-rexp(30,rate)
mean(x)
sqrt(var(x))
qnorm(0.975)

```

- (b). Count the number of such intervals (out of 100) containing the unknown mean (repeated 10 times to see how the coverage rates vary).

```

for(i in 1:10){
y<-0
for(j in 1:100){
x<-rexp(30,rate)
lcl<-mean(x)-qnorm(0.975)*sqrt(var(x))/sqrt(30)
ucl<-mean(x)+qnorm(0.975)*sqrt(var(x))/sqrt(30)
if(lcl<mu && mu<ucl) y<-y+1}
print(y)}

```

How do the coverage rates compare to the expected 95%?

- (c). Count the number of simulation-based intervals (out of 100) containing the unknown mean.

```

y<-0
for(j in 1:100){
x<-mean(rexp(30,rate))
for(i in 1:99){
x<-c(x,mean(rexp(30,rate)))}
if(quantile(x,0.025)<mu && mu<quantile(x,0.975)) y<-y+1 }
print(y)

```

How does the simulation-based method compare to the targeted 95% coverage rate?

**4.4** Repeat Exercise 4.3 using samples of size  $n = 4$  rather than samples of size  $n = 30$ . In all fairness, the Central Limit Theorem cannot be expected to provide valid confidence intervals for such a small sample size; however, the comparison between the “default” confidence interval and the simulation-based interval is of interest here. Note that the percentiles of the  $t$ -distribution are being used here. Even though the sampled population is not Normal, the use of  $t$ -percentiles gives a wider interval than the use of  $z$ -percentiles. This has the effect of making the intervals more likely to contain the unknown mean. No  $t$ - or  $z$ -percentiles are needed for the simulation approach.

- (a). Calculate a single 95% confidence interval for the unknown mean.

```
mu<-runif(1,0.25,2.25)
rate<-1/mu
x<-rexp(4,rate)
mean(x)
sqrt(var(x))
qt(0.975,3)
```

- (b). Count the number of such intervals (out of 100) containing the unknown mean (repeated 10 times to see how the coverage rates vary).

```
for(i in 1:10){
y<-0
for(j in 1:100){
x<-rexp(4,rate)
lcl<-mean(x)-qt(0.975,3)*sqrt(var(x))/sqrt(4)
ucl<-mean(x)+qt(0.975,3)*sqrt(var(x))/sqrt(4)
if(lcl<mu && mu<ucl) y<-y+1}
print(y)}
```

How do the coverage rates compare to the targeted 95%?

- (c). Count the number of simulation-based intervals (out of 100) containing the unknown mean.

```
y<-0
for(j in 1:100){
x<-mean(rexp(4,rate))
for(i in 1:99){
x<-c(x,mean(rexp(4,rate)))}
if(quantile(x,0.025)<mu && mu<quantile(x,0.975)) y<-y+1 }
print(y)
```

How does the simulation-based method compare to the targeted 95% coverage rate?

**4.5** The length of stay in days (LOS) of patients with digestive system disorders admitted to a particular hospital during a specific year is the population of interest. The data is in the file `drg183.txt` which is available via ftp from a VINCENT prompt as follows:

```
cd stat415
ftp likelihood.stat.iastate.edu
anonymous.wmd
youremail@iastate.edu
get drg183.txt
quit
```

To get the data into an S object, start S and type the command

```
drg183<-scan("drg183.txt")
```

Type `drg183` to see the data.

This data set has 404 observations. It is the *population of interest* for this exercise! Keep in mind that one usually does not have the population data at hand. The population is made available here because we want to compare the results of various estimation methods.

For this exercise, we want to obtain a 95% confidence interval for the mean (yes, the mean) LOS for digestive system disorder patients, so  $\theta = \mu$ . However, we only have LOS values for a sample of  $n = 18$  such patients. The usual procedure is to use the standard  $t$ -percentile-based confidence interval, but this is only valid when the population is Normally-distributed.

The following S+ commands will generate four confidence intervals for  $\mu$  using the methods: standard  $t$ -interval, percentile method,  $BC_a$  method, and bootstrap- $t$  method, respectively. After executing each block of S+ commands, record the confidence interval. *After* you have obtained all four intervals, then type `mean(drg183)` to obtain the “unknown” value of  $\theta = \mu$ . Record this value and note whether or not each of your intervals contained the value.

```
x<-sample(drg183,18)
xbar<-mean(x)
s<-sqrt(var(x))
t<-qt(0.975,17)
lcl<-xbar-t*s/sqrt(18)
ucl<-xbar+t*s/sqrt(18)
lcl
ucl

y<-bootstrap(x,1000,mean)
thetastars<-y[[1]]
lcl<-quantile(thetastars,0.025)
ucl<-quantile(thetastars,0.975)
lcl
ucl

y<-bcanon(x,1000,mean)
lcl<-y[[1]][1,2]
ucl<-y[[1]][8,2]
lcl
ucl

y<-boott(x,mean)
lcl<-y[[1]][1,3]
ucl<-y[[1]][1,9]
lcl
ucl
```

**4.6** See Exercise 4.5 for a description of the data to be used in this exercise. In Exercise 4.5, you were instructed to obtain one confidence interval for  $\mu$  by various methods. In this exercise, you will evaluate the coverage rates of the various methods.

The following S+ commands generate one hundred standard  $t$ -intervals (10 times, so that you can see how coverage rates vary).

```
for(i in 1:10){
```

```

count<-0
for(j in 1:100){
x<-sample(drg183,18)
xbar<-mean(x)
s<-sqrt(var(x))
t<-qt(0.975,17)
lcl<-xbar-t*s/sqrt(18)
ucl<-xbar+t*s/sqrt(18)
mu<-mean(drg183)
if(lcl<mu && mu<ucl) count<-count+1}
print(count)}

```

How do the coverage rates compare to the target 95% rate? Type `lcl` and `ucl` and record the values—this is the last confidence interval that was calculated.

Use the following command to get a histogram of the population (something you can't do in the “real world”). I've chosen 16 classes because the default of `S+` is not too informative in this case.

```
hist(drg183,nclass=16)
```

Does the population appear to be Normal? What does this imply about using the standard  $t$ -interval? (Note: with a small sample like 18, the Central Limit Theorem is not applicable, so the standard  $z$ -interval is not an option.)

The following `S+` commands will generate 100 95% bootstrap confidence intervals (percentile method) and count how many contain  $\mu$ . Here, the 100 intervals are only generated once because repeating the loop 10 times may take some time!

```

count<-0
for(j in 1:100){
x<-sample(drg183,18)
y<-bootstrap(x,1000,mean)
thetastars<-y[[1]]
lcl<-quantile(thetastars,0.025)
ucl<-quantile(thetastars,0.975)
mu<-mean(drg183)
if(lcl<mu && mu<ucl) count<-count+1}
print(count)

```

How does the coverage rate compare with the target 95% rate? Type `lcl` and `ucl` and record the values—this is the last confidence interval that was calculated.

The following `S+` commands will generate 100 95% bootstrap confidence intervals ( $BC_a$  method) and count how many contain  $\mu$ . Here, the 100 intervals are only generated once because repeating the loop 10 times may take some time!

```

count<-0
for(j in 1:100){
x<-sample(drg183,18)
y<-bcanon(x,1000,mean)
lcl<-y[[1]][1,2]
ucl<-y[[1]][8,2]
mu<-mean(drg183)
if(lcl<mu && mu<ucl) count<-count+1}
print(count)

```

How does the coverage rate compare with the target 95% rate? Type `lc1` and `uc1` and record the values—this is the last confidence interval that was calculated.

The following S+ commands will generate 100 95% bootstrap confidence intervals (Bootstrap- $t$  method) and count how many contain  $\mu$ . Here, the 100 intervals are only generated once because repeating the loop 10 times may take some time!

```
count<-0
for(j in 1:100){
  x<-sample(drg183,18)
  y<-boott(x,mean)
  lc1<-y[[1]][1,3]
  uc1<-y[[1]][1,9]
  mu<-mean(drg183)
  if(lc1<mu && mu<uc1) count<-count+1}
print(count)
```

How does the coverage rate compare with the target 95% rate? Type `lc1` and `uc1` and record the values—this is the last confidence interval that was calculated.

Summarize your findings. Also, type `mu` and record the value of  $\theta = \mu$  (this should be the same value you obtained in Exercise 4.5). Which of your four confidence intervals contained this value?