Chapter 14
Replication of Results

We have covered statistical theory, probability, sampling distributions, statistical distributions, hypothesis testing, and various statistical tests based on the level of measurement and type of research design. Research for many decades involved carrying out a single study, i.e., collection of a single random sample of data from the population. Researchers today are becoming more concerned with replicating their research results. However, time and resources often do not permit conducting a research study again. Researchers instead have created techniques that provide some level of replicating their findings.

Every day we ask ourselves important questions. It could be as simple as what route to take to a new job. In the process of answering the question, we gather information or data. This could include asking co-workers, driving different routes, and examining an area roadmap. Once we feel that sufficient information has been collected, we answer the question for ourselves and often share the answer with others. Trying the route to work we selected validates our findings or conclusion. The process of asking a question, gathering data, answering the question, and validating the conclusion is the key to the research process.

Once we have asked an important question, gathered data, and answered the question, others may ask whether the same results would occur if the process were repeated. In other words, could the research findings be replicated? In order to repeat the same process, we must first document the methods and procedures used in the original research. Then, another person can replicate the research process and report their findings. If the research findings are valid and consistent, then others should report findings similar to the original research.

In numerous academic disciplines, research findings are reported along with the methods and procedures used in the study. Unfortunately, not many research studies are replicated. This is due primarily to the time, money, and resources needed to replicate research studies. Instead, other approaches have been developed which don't require conducting the study again. These methods include cross-validation, jackknife, and bootstrap.

R. Schumacker and S. Tomek, *Understanding Statistics Using R*,
DOI 10.1007/978-1-4614-6227-9_14, © Springer Science+Business Media New York 2013

Cross Validation

The cross-validation approach involves taking a random sample of data from a population. Typically we would compute a sample statistic for the sample data of size N where the sample statistic is our estimator of the population parameter. In the **cross-validation** approach, the original sample data are randomly split into two equal halves. A sample statistic is computed using one half of the sample data and applied to the other half of the sample data. If the sample statistics for the two randomly split data halves are similar, we assume the research findings would be replicable. Otherwise, the findings are not consistent and probably could not be replicated. A large random sample is generally needed to provide two randomly split data halves that are of sufficient sample size so as not to affect the sample statistics. For example, a sample of size N = 1000 randomly drawn from a population would be randomly split into equal halves; each cross-validation sample size would be N = 500.

The essence of the cross-validation technique is the computing of a sample statistic on the first set of sample data and applying it to the second set of sample data. This technique is different from other approaches in statistics. We are not comparing sample means from two random samples of data drawn from different populations and testing whether the means are similar or dissimilar (independent t-test for mean differences). We are not testing two sample proportions to determine if they are similar or dissimilar (z-test for differences in proportions). We are not comparing a sample mean to a population mean (one sample t-test). It should be noted that if the original sample statistic was not a good estimator of the population parameter, the cross-validation results would not improve the estimation of the population parameter.

Replicating research results involves conducting another study using the same methods and procedures. When unable to replicate another study, researchers apply the cross-validation approach to a single sample of data randomly drawn from the population. The cross-validation approach involves randomly splitting a random sample into two equal halves, then computing a sample statistic on one sub-sample and applying it to the other sub-sample. The cross-validation approach is not the same as other statistical tests, which randomly sample from a population and test sample estimates of the population parameter. The cross-validation approach does not improve the sample statistic as an estimate of the population parameter. A comparison of two regression equations from two equal halves of sample data will indicate the stability of the intercept and regression weight.

CROSS VALIDATION Programs

The cross validation programs take a random sample from a population. The sample data will be randomly split into two equal halves of size, N/2. The **linear regression** equation, $Y = a + bX + e$, will be computed using the first one-half sample. The regression equation will then be applied to the second one-half sample. The regression weights and R-square values will be compared. A smaller R-squared value in

the second one-half sample is expected because the regression equation coefficients were selected in the first one-half sample data to minimize the sum of squared errors (least squares criterion). The lower R-squared value in the second data set is referred to as the *shrinkage* of the R-squared value. The R-squared value in the second one-half data set will not always be lower than the R-squared value in the first one-half data set. The second cross validation program yields results to determine the amount of expected shrinkage in the second one-half sample regression statistics.

The first program, **CROSSVALIDATION1** uses a traditional method of cross validation whereby a regression equation is created on one randomly chosen half of a sample and applied to the data in the second half to compare stability of regression weights and check for R-squared shrinkage. The second program, **CROSSVALIDATION2**, creates separate independent regression equations on each randomly split half and compares the regression weights and R-squared values. Both programs begin by assigning values for the intercept and slope of the true population regression equation, the sample size, and the number of replications. A replication in the program refers to the splitting of the sample size into random halves and applying the particular method of cross validation. For each replication, a full sample is split into two new random halves, so the same full sample from the population is used for all replications within the same run of the program.

The regression coefficients are computed using the **lsfit** function and the results assigned to the *sampleReg* object. The **ls.print** function is used to output the regression summary statistics and permits selection of specific results, for example, the R-squared value for the full sample is assigned to *sampleR2* and the regression weight is assigned to b*Sample*. The other half of the sample is assigned to vectors to be analyzed using the vector notation of [−*halfPoints*], which means all points that were not in the first vector of random values. The predicted Y-values are then determined from the X-values of the second half using the slope and intercept determined by the regression of the first half of the sample, the sum of squared errors for the regression, the total sum of squared error Y-values and the predicted Y-values. The regression weight and R-squared value for the full sample is displayed using the **cat** function and the output results given by the **print** function. The second program is identical to the first, except that the second half of the sample data is analyzed separately with the **lsfit** function. The output of the second program therefore includes separate regression results for both samples of data for comparison.

CROSS VALIDATION Program Output

CROSSVALIDATION1

```
Population

Regression equation: Y = 3 + 0.25
Sample Size=1000
N Replications=5
```

```
Full Sample

Regression equation: Y = 2.93 + 0.262
R-Squared=0.3123

              Sample A     Sample A    Sample B
              Reg Weight   R-Squared   R-Squared
Replication 1 0.271        0.3476      0.324
Replication 2 0.288        0.3615      0.399
Replication 3 0.255        0.3041      0.289
Replication 4 0.279        0.338       0.373
Replication 5 0.265        0.3124      0.331
```

CROSSVALIDATION2

```
Population

  Regression equation: Y = 3 + 0.25 (X)
  Sample Size=1000
  N Replications=5

  Sample

  Regression equation: Y=3.086 + 0.241 (X)
  R-Squared=0.2735

               Sample A    Sample A    Sample B    Sample B
               Reg Weight  R-Squared   Reg Weight  R-Squared
Replication 1  0.243       0.2775      0.24        0.2725
Replication 2  0.238       0.2726      0.243       0.2742
Replication 3  0.264       0.3172      0.218       0.2318
Replication 4  0.247       0.2745      0.234       0.2724
Replication 5  0.253       0.2717      0.229       0.2767
```

Cross Validation Exercises

1. Run the CROSSVALIDATION1 program 5 times for an original sample size of 500. Record the regression weight and the two R-square values for *each* sample of size N = 250. Record the regression weight and R-square value for the original sample.

   a. Original Sample (N = 500): Regression Weight _____
   R-Square _____.

b.

	Sample A (N = 250) Regression weight	R-Square	Sample B (N = 250) R-Square
1.	_____	_____	_____
2.	_____	_____	_____
3.	_____	_____	_____
4.	_____	_____	_____
5.	_____	_____	_____

c. The regression equation computed for sample A is applied to sample B. Are the R-squared values similar in sample A and sample B for the five replications?

YES _____NO _____.

2. Run the CROSSVALIDATION1 program with 5 replications and a sample size of 1,000.

a. Compare the regression equation results for the original samples of N = 500 and N = 1000. Record the SAMPLE regression weights and R-square values.
N = 500 Regression Weight_____ R-square _____
N = 1000 Regression Weight_____ R-square _____
b. Does sample size affect the R-square value? YES _____NO _____.
c. Which sample size would give better estimates of R-square and the regression weight?
N = 500 _____ N = 1000 _____

3. Run the CROSSVALIDATION2 program with 5 replications using an original sample size of 500. Record the regression weights and R-square values for *each* sample of size N = 250. Record the regression weight and R-square value for the original sample.

a. Sample (N = 500): Regression Weight _____ R-Square _____.
b.

	Sample A (N = 250) Regression Weight	R-Square	Sample B (N = 250) Regression Weight	R-Square
1.	_____	_____	_____	_____
2.	_____	_____	_____	_____
3.	_____	_____	_____	_____
4.	_____	_____	_____	_____
5.	_____	_____	_____	_____

c. The regression equations for sample A and sample B are computed independently. Are the R-squared values similar in sample A and sample B for the five replications?

YES _____NO _____.

4. Run the CROSSVALIDATION2 program with 5 replications having an original sample size of 1,000.

a. Compare the regression equation results for the original samples of N = 500 and N = 1000. Record the SAMPLE regression weights and R-square values.
N = 500 Regression Weight_____ R-square _____
N = 1000 Regression Weight_____ R-square _____
b. Does sample size affect the R-square value? YES _____NO _____.
c. Which sample size would give better estimates of R-square and the regression weight? N = 500 _____ N = 1000 _____

Jackknife

The jackknife procedure involves the use of a single random sample of data drawn from a population of data. Recall that the sample statistic is an estimate of the population parameter. We also learned in previous chapters that how good the sample statistic is as an estimate of the population parameter depends on the sample size. The jackknife procedure is concerned with whether the sample statistic as an estimate of the population parameter is affected by any single data value. For example, we know that the sample mean is affected by extreme data values, which is indicated by the standard deviation of the sample data.

The jackknife approach uses a single sample of data, computes the original sample statistic (e.g., sample mean), and then computes the sample statistic for each sample of size N − 1. Basically, each sample mean after the original sample mean would be computed based on the omission of one data point. The jackknife approach is therefore useful in determining whether an influential data point exists that dramatically changes the sample statistic. The jackknife procedure can be applied to any sample statistic based on a random sample drawn from a population. The jackknife method computes a jackknife mean based on the exclusion of a different data point each time. The number of jackknife replications therefore typically equals the original sample size so that the influence of each data point on the sample statistic can be determined.

An example might help to better understand the jackknife procedure. A random sample of 10 numbers is drawn from a population. The numbers are 2, 4, 9, 12, 8, 7, 15, 11, 3, and 14. The sum of the numbers is 85. The mean is calculated as 85 divided by 10, which equals 8.5. This sample mean is an estimate of the population mean. The jackknife procedure calculates 10 additional sample means based on N = 9, but each time with a different data value omitted. To compute each jackknife

mean, the data value omitted is subtracted from the sum of 85 and this new sum is divided by N = 9 to yield the jackknife mean. The jackknife procedure for these data values is outlined below.

Sample Size	Mean	Values Used	Value Omitted
10	8.5	2, 4, 9, 12, 8, 7, 15, 11, 3, 14	None
9	9.2	4, 9, 12, 8, 7, 15, 11, 3, 14	2
9	9.0	2, 9, 12, 8, 7, 15, 11, 3, 14	4
9	8.4	2, 4, 12, 8, 7, 15, 11, 3, 14	9
9	8.1	2, 4, 9, 8, 7, 15, 11, 3, 14	12
9	8.5	2, 4, 9, 12, 7, 15, 11, 3, 14	8
9	8.6	2, 4, 9, 12, 8, 15, 11, 3, 14	7
9	7.7	2, 4, 9, 12, 8, 7, 11, 3, 14	15
9	8.2	2, 4, 9, 12, 8, 7, 15, 3, 14	11
9	9.1	2, 4, 9, 12, 8, 7, 15, 11, 14	3
9	7.8	2, 4, 9, 12, 8, 7, 15, 11, 3	14

The jackknife sample means ranged from 7.7 to 9.2 with the original sample mean of 8.5. The omission of a low data value inflated (increased) the sample mean as an estimate of the population mean. The omission of a high data value deflated (lowered) the sample mean as an estimate of the population mean. Both of these outcomes are expected and help us to understand the nature of how extreme data values (outliers) affect the sample statistic as an estimate of a population parameter.

Another example of the jackknife procedure will highlight the detection of an influential (outlier) data value. Once again, we randomly sample 10 data values from a population. The sum of the numbers is 158 with an original sample mean of 15.8. The results are summarized below. The jackknife means in this second example ranged from 9.2 to 17.2 with an original sample mean of 15.8. Notice that the original sample mean of 15.8 is less than every other jackknife mean, except the one with an omitted influential data value, i.e., 75. The results indicate how the removal of an influential data value can increase or decrease the sample statistic value.

Sample Size	Mean	Values Used	Value Omitted
10	15.8	75, 4, 9, 12, 8, 7, 15, 11, 3, 14	None
9	9.2	4, 9, 12, 8, 7, 15, 11, 3, 14	75
9	17.1	75, 9, 12, 8, 7, 15, 11, 3, 14	4
9	16.5	75, 4, 12, 8, 7, 15, 11, 3, 14	9
9	16.2	75, 4, 9, 8, 7, 15, 11, 3, 14	12
9	16.6	75, 4, 9, 12, 7, 15, 11, 3, 14	8
9	16.7	75, 4, 9, 12, 8, 15, 11, 3, 14	7
9	15.9	75, 4, 9, 12, 8, 7, 11, 3, 14	15
9	16.3	75, 4, 9, 12, 8, 7, 15, 3, 14	11
9	17.2	75, 4, 9, 12, 8, 7, 15, 11, 14	3
9	16.0	75, 4, 9, 12, 8, 7, 15, 11, 3	14

An important comparison can be made between the first and second example. In the first example, the original sample mean of 8.5 fell between all jackknife means, which ranged from 7.7 to 9.2. In the second example, the jackknife means were all greater than the original sample mean, with the exception of the one influential data value. If we examine the jackknife means for each omitted data value, it points out the presence of an influential data value (outlier or extreme value).

The descriptive information for the jackknife means in both examples also clearly indicates the presence of influential data in the second example. The jackknife means were treated as new data such that the average is the *mean* of the jackknife means. The standard deviation, variance, and 95% confidence interval indicate more data dispersion in the second example. This dispersion is interpreted as less accuracy, more variability, and greater difference between the original sample mean and the jackknife means. The descriptive information is:

Descriptive Information	First Example	Second Example
Sample Size	N = 10	N = 10
Range	7.7–9.2	9.2–17.2
Mean	8.46	15.77
Standard Deviation	.52	2.35
Variance	.27	5.52
95% Confidence Interval	(7.44, 9.48)	(11.16, 20.38)

The jackknife procedure uses the original random sample drawn from a population to estimate additional sample means based on N − 1 sample data points. The jackknife procedure is useful for identifying influential, extreme, or outlier data values in a random sample of data. The jackknife method can be used with any sample statistic that is computed from a random sample of data drawn from a well-defined population. The jackknife approach helps to validate whether a sample of data provides a good sample statistic as an estimator of the population parameter. The number of jackknife means is equal to the sample size for the purposes of detecting an influential data value. The confidence interval around a set of jackknife means will be smaller when influential data values are not present.

JACKKNIFE R Program

The JACKKNIFE program utilizes the jackknife function in the R library bootstrap. Therefore the program must first issue the command: library(bootstrap). The jackknife function then permits an easy method for calculating the mean and percentiles after jackknifing the samples. The calculations could be done manually instead of using the jackknife function, but the function makes it easier to compute for large samples. The jackknife function is given the vector of data values to be jackknifed. The mean and standard deviation for the entire sample are printed out, followed by the summary of the jackknife results.

JACKKNIFE Program Output

```
   Data Values=10 20 30 40 50 60 70 80 90 100

   Original mean=55 Original standard deviation=30.28

Sample Size  Jackknife Mean  Values Used                 Value Omitted
9            60              20,30,40,50,60,70,80,90,100 10
9            58.89           10,30,40,50,60,70,80,90,100 20
9            57.78           10,20,40,50,60,70,80,90,100 30
9            56.67           10,20,30,50,60,70,80,90,100 40
9            55.56           10,20,30,40,60,70,80,90,100 50
9            54.44           10,20,30,40,50,70,80,90,100 60
9            53.33           10,20,30,40,50,60,80,90,100 70
9            52.22           10,20,30,40,50,60,70,90,100 80
9            51.11           10,20,30,40,50,60,70,80,100 90
9            50              10,20,30,40,50,60,70,80,90  100
```

Jackknife Exercises

1. Run the JACKKNIFE program with the following 10 data values as a random Sample A: data<– c(1,2,3,4,5,6,7,8,9,10). Print the Graph.

   a. Record the Original mean and standard deviation.
   Original Mean=_____ Original Standard Deviation=_____

   b. Record the following information for the N – 1 data sets.

		Jackknife	
Run	Sample Size	Mean	Value Omitted
1	9		
2	9		
3	9		
4	9		
5	9		
6	9		
7	9		
8	9		
9	9		
10	9		

   c. Calculate the Standard Error of the Mean using Original sample standard deviation and the number of replications, N=10. $SE = S/\sqrt{N}$ = _____=_____.

d. Calculate the 95% Confidence Interval using the Original sample mean and standard deviation. $95\%CI = \overline{X} \pm 1.96(S)$ = (_____, _____)

e. List the Range of Jackknife Means. Highest Mean _____ Lowest Mean _____

f. How many Jackknife means are higher than the Original Sample mean? _____

2. Run the JACKKNIFE program with the following 10 data values as a random Sample B: data <- c(1,2,3,4,5,6,7,8,9,100). Print the Graph.

a. Record the Original sample mean and standard deviation.
Original Mean = _____ Original Standard Deviation = _____

b. Record the following information for the N − 1 data sets

		Jackknife	
Run	Sample Size	Mean	Value Omitted
1	9		
2	9		
3	9		
4	9		
5	9		
6	9		
7	9		
8	9		
9	9		
10	9		

c. Calculate the Standard Error of the Mean using Original sample standard deviation and the number of replications, N = 10. $SE = S/\sqrt{N}$ = _____ = _____.

d. Calculate the 95% Confidence Interval using the Original sample mean and standard deviation. $95\%CI = \overline{X} \pm 1.96(S)$ = (_____, _____)

e. List the Range of Jackknife Means. Highest Mean _____ Lowest Mean _____

f. How many Jackknife means are higher than the Original Sample mean? _____

3. List the Original sample means and standard deviations, SE, 95% CI, and range of Jackknife means for Sample A and Sample B above.

	Sample Mean	Sample SD	SE	95%CI	Jackknife Range
Sample A	_____	_____	_____	(_____, _____)	High: _____ Low: _____
Sample B	_____	_____	_____	(_____, _____)	High: _____ Low: _____

a. Does Sample A or Sample B have a higher sample mean?
   Sample A _____ Sample B _____
b. Does Sample A or Sample B have a larger standard deviation?
   Sample A _____ Sample B _____
c. Does Sample A or Sample B have a larger Standard Error of the Mean?
   Sample A _____ Sample B _____
d. Does Sample A or Sample B have a wider 95% Confidence Interval?
   Sample A _____ Sample B _____
e. Does Sample A or Sample B have more Jackknife Means higher than the Original sample mean?
   Sample A _____ Sample B _____
f. Which sample has a more accurate sample mean estimator of the population mean?
   Sample A _____ Sample B _____
g. Summarize **a** to **e** above in regard to their indicating influential data points.

_____

_____

_____

Bootstrap

The bootstrap method differs from the traditional parametric approach to inferential statistics because it uses *sampling with replacement* to create the sampling distribution of a statistic. The bootstrap method doesn't take a random sample from a population in the same way as that used in our previous inferential statistics. The bootstrap method is useful for reproducing the sampling distribution of any statistic, e.g., the median or regression weight. The basic idea is that conclusions are made about a population parameter from a random sample of data, but in which a sampling distribution of the statistic is generated based on sampling with replacement. The factors that influence the shape of the sampling distribution are therefore important because it is the bootstrap estimate from the sampling distribution that allows us to make an inference to the population.

The bootstrap procedure uses a random sample of data as a substitute for the population data. The randomly sampled data acts as a "pseudo" population from which the bootstrap method repeatedly samples the data. The repeated sampling of data is done with replacement of each data point after selection. The resampling technique therefore samples from the "pseudo" population using the same set of data values each time. Since each data value is replaced before taking the next random selection, it is possible to have the same data value selected more than once and used in the calculation of the final sample statistic. The probabilities in the earlier chapters of the book were based on randomly sampling *without* replacement where each individual, object, or event had an equally likely chance of being selected.

The bootstrap method uses probabilities based upon randomly sampling *with* replacement where each individual, object, or event has an equally likely chance of being selected each time a data value is randomly drawn.

The bootstrap procedure involves the following steps:

Step 1: A random sample of data for a given sample size N is drawn from the population with mean, μ, and standard deviation, σ.

Step 2: The random sample of data size N acts as a "pseudo" population with mean, μ*, and standard deviation, σ*.

Step 3: The bootstrap method takes *n* bootstrap samples of sample size N from the "pseudo" population, each time replacing the randomly sampled data point. For each sample of size N a sample statistic is computed.

Step 4: A frequency distribution of the *n* bootstrap sample statistics is graphed which represents the sampling distribution of the statistic. The mean of this sampling distribution is the bootstrap estimate, θ*, which has a standard error of SE_{θ^*} computed by:

$$SE_{\theta} = \sqrt{\frac{\sum(\theta_i^* - \theta^*)^2}{n-1}}$$

Step 5: The amount of bias in the original sample statistic as an estimate of the population parameter is calculated by subtracting: $\mu^* - \theta^*$. If the bootstrap estimate is similar to the corresponding "pseudo" population parameter, then no bias is present. A small difference would still indicate that the original sample statistic is a good estimator of the population parameter.

Step 6: Calculate a confidence interval around the bootstrap estimate using the standard error of the bootstrap estimate and level of significance, Z. The confidence interval is computed by:

$$\mathrm{CI}_{\theta} = \theta^* \pm \mathrm{Z}(\mathrm{SE}_{\theta^*}).$$

The bootstrap method can be based on samples of size N that equal the original sample size N or are larger when it involves sampling data points with replacement. Most applications resample to produce sample sizes equal to the "pseudo" population size. The bootstrap method can also be used to determine the amount of bias between any sample statistic and population parameter. It should be noted that the bootstrap method is only useful for determining the amount of bias in the sample statistic when the original sample is randomly drawn and representative of the population. This makes sense because if the original sample data were not representative of the population, then creating the sampling distribution of this data would erroneously indicate population characteristics.

An example will help to clarify how the bootstrap method is used to determine the amount of bias in the original sample statistic as an estimate of the population parameter. A random sample of 100 data points is drawn from the population. This random sample now becomes a "pseudo" population. The "pseudo" population has a mean, $\mu^* = 50$, and standard deviation, $\sigma^* = 20$. The bootstrap procedure will randomly sample data points with replacement from this "pseudo" population. The bootstrap sample size will be $n = 10$ and the number of bootstrap samples will be $N = 5$. The resulting data for each bootstrap sample is given below.

Sample Run	Size	Bootstrap Data	Bootstrap Mean	Bootstrap Standard Deviation
1	10	10, 30, 35, 40, 50, 80, 90, 75, 20, 60	49.0	26.75
2	10	15, 25, 75, 40, 55, 55, 95, 70, 30, 65	52.5	24.97
3	10	85, 45, 35, 25, 45, 60, 75, 80, 90, 15	55.5	26.40
4	10	20, 30, 45, 50, 55, 65, 10, 70, 85, 95	52.5	27.41
5	10	50, 50, 20, 45, 65, 75, 30, 80, 70, 30	51.5	20.69

The bootstrap estimate, based on the average of the sampling distribution of bootstrap means is $\theta^* = 52.2$. The standard error of the bootstrap estimate, based on the square root of the sum of squares difference between each bootstrap sample mean and the bootstrap estimate, divided by the number of bootstrap samples minus one, is $SE_{\theta^*} = 2.33$. To establish a 95% confidence interval, a $Z = 1.96$ value under the normal distribution is used. The 95% confidence interval is therefore computed as:

$$95\%\ CI_{\theta} = \theta^* \pm Z\ (SE_{\theta^*})$$
$$95\%\ CI_{\theta} = 52.2 \pm 1.96\ \%21.80 / 4)$$
$$95\%\ CI_{\theta} = 52.2 \pm 1.96\ (2.33)$$
$$95\%\ CI_{\theta} = 52.2 \pm 4.57$$
$$95\%\ CI_{\theta} = (47.63, 56.77)$$

The amount of bias is indicated by, $\mu^* - \theta^*$, which is $50 - 52.2 = -2.2$. On the average, the bootstrap means were 2.2 units higher than the "pseudo" population mean. The sign of the *bootstrap estimate* will be either positive or negative depending upon whether the bootstrap estimate is lower or higher than the "pseudo" population parameter, respectively. Since the *bootstrap confidence interval* captures the "pseudo" population mean, we would conclude that the original sample mean is a good stable estimate of the population mean.

The bootstrap method uses a random sample of data as a "pseudo" population. The bootstrap procedure resamples the "pseudo" population with replacement. The bootstrap samples of data can contain some of the same data points. The bootstrap estimate is the average of the bootstrap sample statistics. The bootstrap standard deviation is based on the bootstrap data. The bootstrap confidence interval should capture the "pseudo" population parameter when the original sample statistic is a good estimate of the population parameter. The bootstrap method is used to determine the amount of bias between the "pseudo" population parameter and the bootstrap estimate. Similar values indicate no bias. The bootstrap method can be used

with any sample statistic to help determine if the sample statistic is a good or stable estimator of the population parameter. The bootstrap method is not useful when the random sample of data is not representative of the population data.

BOOTSTRAP R Program

The BOOTSTRAP program uses the built-in **bootstrap** function from library(bootstrap). The program inputs the sample size and the number of bootstrap samples. A random sample of data from a normal distribution with a mean of 50 and standard deviation of 10 is generated. The bootstrap function is then performed for the number of bootstrap samples when calculating the mean. The Observed mean corresponds to the mean of the sample from the population and the Bootstrap Mean corresponds to the mean of the bootstrap samples. The Bias is the difference between the Observed Mean and the Bootstrap Mean. The standard error of the bootstrap estimates is reported and used to create the 95% confidence interval around the Bootstrap Mean.

The Observed Mean falls within the 95% confidence interval when the sample data is considered representative of the population.

BOOTSTRAP Program Output

```
Sample Size=100
N Bootstraps=500

Observed Mean = 52.51114
Bootstrap Mean=50.95276

Bias=1.558382
SE=2.057323

95% CI=Bootstrap Mean +/- 1.96SE
95% CI=( 48.89543 53.01008 )
```

Bootstrap Exercises

1. Run the BOOTSTRAP program for a random sample of N = 200, then take 20 bootstrap samples. The program settings should be:

```
sampleSize<- 200
numBootstraps<- 20
```

a. Record the Observed Mean, Bootstrap Mean, Bias, and Standard Error.
   Observed Mean _____
   Bootstrap Mean _____
   Bias _____
   SE _____
b. Calculate the 95% Confidence Interval around the Bootstrap Mean.
   95%CI = Bootstrap Mean +/– 1.96 (SE) = (_____, _____)
c. What does the bootstrap results indicate given the bias and confidence interval?

   _____

   _____

   _____

2. Run the BOOTSTRAP program for a random sample of N = 1000, then take 40 bootstrap samples. The program settings should be:

```
sampleSize <- 1000
numBootstraps <- 40
```

a. Record the Observed Mean, Bootstrap Mean, Bias, and Standard Error.
   Observed Mean _____
   Bootstrap Mean _____
   Bias _____
   SE _____
b. Calculate the 95% Confidence Interval around the Bootstrap Mean.

   95% CI = Bootstrap Mean +/– 1.96 (SE) = (_____, _____)
c. What does the bootstrap indicate given the bias and confidence interval?

   _____

   _____

   _____

d. Does the number of bootstrap samples provide a better bootstrap estimate?
   YES ____ NO _____
e. What would happen if the random sample of data from the population was not representative?

   _____

True or False Questions

Cross Validation

T F a. Cross-validation techniques verify that the sample statistic is a good estimator of the population parameter.
T F b. Cross-validation techniques involve splitting a random sample from a population into two equal halves.
T F c. The cross-validation approach involves computing sample statistics using one sub-sample and applying them to the second sub-sample.
T F d. Cross-validation techniques require large sample sizes.
T F e. A replication of findings generally requires conducting another study using the methods and procedures of the original study.

Jackknife

T F a. The jackknife procedure is useful for detecting influential data values.
T F b. A jackknife mean computed with an influential data value does not fall within the confidence interval of the jackknife means.
T F c. The standard deviation of the jackknife means shows whether more variability is present, hence influential data values.
T F d. The jackknife approach can be used with any sample statistic computed from random samples of data drawn from a population.
T F e. The jackknife procedure repeatedly samples the population data to determine if the sample mean is a good estimate of the population mean.
T F f. The number of jackknife means is typically equal to the original sample size for the purposes of detecting influential data points.

Bootstrap

T F a. The bootstrap method is used to determine if the sample statistic is a stable estimator of the population parameter.

T F b. The "Observed Mean" will always fall in the bootstrap confidence interval.

T F c. A random sample of data must be representative of the population before the bootstrap procedure is accurate.

T F d. The bootstrap procedure involves sampling data with replacement.

T F e. No bias between the "observed mean" and the bootstrap mean estimate indicates that the sample statistic is a good estimator of the population parameter.

T F f. The bootstrap method can be used with any sample statistic computed from a random sample of data.

T F g. The bootstrap procedure creates random samples of data where each data value is unique.

T F h. The resampling method draws random samples from the "pseudo" population using the same set of data values each time.