Chapter 5
Central Tendency and Dispersion

Central Tendency

In the previous chapter, unimodal and bimodal score distributions were demonstrated. In addition to the mode or most frequent score in a sample of data, the mean and median are also considered measures of central tendency. **Central tendency** is where most scores occur in the middle of a symmetrical distribution and then spread out. The mode, mean, and median values will all be identical in a normal distribution.

This chapter examines the effect upon means and medians when data values are transformed and/or extreme data values are added to a data set. The **mean** score is the arithmetic average of numbers in a data set. The mean is computed by taking the sum of the numbers and dividing by the total. The **median** score is the middle score found by arranging a set of numbers from the smallest to the largest (or from the largest to the smallest). If the data set contains an odd number of values, the median is the middle value in the ordered data set. If there is an even number of data values, the median is the average of the two middle values in the ordered data set. The median value is the score that divides the distribution into two equal halves.

Sample data are sometimes modified or transformed to permit comparisons and aid in the interpretation of sample estimates. For example, if the length of cars in inches was changed to meters, one could multiply 0.0254 times the car length to yield meters. What effect does this multiplication have on the mean value representing length of cars? What if an instructor decides to adjust a set of test grades for an exceptionally long test by adding ten points to each student's score? What effect does this addition have on the mean score? If the price of the most expensive house in a neighborhood increases, what happens to the median value of the houses in that neighborhood? What happens to the average value of houses in that neighborhood? These basic effects can be seen because:

R. Schumacker and S. Tomek, *Understanding Statistics Using R*,
DOI 10.1007/978-1-4614-6227-9_5, © Springer Science+Business Media New York 2013

If a constant is added to all the values in a data set:

- The mean of the modified data is the mean of the initial data set plus the constant.
- The median of the modified data set is the median of the initial data set plus the constant.

If all of the values in a data set are multiplied by a constant:

- The mean of the modified data set is the mean of the initial data set times the constant.
- The median of the modified data set is the median of the initial data set times the constant.

If the largest value in a data set is replaced by a smaller value, then the mean of the modified data set is smaller than the mean of the initial data set, but the median is unchanged. Extreme values affect the mean, but do not affect the median. If the mean of a data set is subtracted from each of the data values, then the mean of the modified data set is 0. The mean and median of the initial data set can be recovered from the mean and median of the modified data set by adding the negative of the added constant or by dividing by the number that was used as a multiplier.

MEAN-MEDIAN R Program

In the MEAN-MEDIAN R program, you will enter an initial data set. The initial data set will have six numbers and the numbers must be entered in order from smallest to largest. This initial data will then be transformed in three different ways: (1) add a constant to each data value; (2) multiply each data value by a constant; or (3) replace the largest data value by a smaller data value. The program will print the initial data set with its mean and median followed by the mean and median from the modified data sets. The effect of these data modifications on the mean and median can then be observed.

MEAN-MEDIAN Program Output

```
Initial Data Set
   2 4 5 9 15 19
   Mean = 9 Median = 7

Added 10 to the Initial data
   Added Value Data
   Mean = 19 Median = 17

Multiplied 5 to the Initial data
   Multiplied Value Data
   Mean = 45 Median = 35
```

```
Replaced Largest Value 19 in the Initial data with 1
   Replaced Value Data
   Mean = 6 Median = 4.5
```

MEAN-MEDIAN Exercises

1. To run MEAN-MEDIAN program, the initial data set (2, 4, 5, 9, 15, 19) is specified.
   Use the following data transformations and record the results.

INITIAL DATA	MEAN	MEDIAN
ADD		
3	_____	_____
−2	_____	_____
10	_____	_____
5	_____	_____
MULTIPLY BY		
2	_____	_____
−10	_____	_____
0.5	_____	_____
REPLACE LAST VALUE WITH		
1	_____	_____
5	_____	_____
10	_____	_____

2. Complete the following statements:

   a. If a constant number is added to each of the data values in a set, then the mean of the modified data is equal to the initial mean _____ and the median of the modified data set is equal to the initial median _____

   _____.

   b. If each data value in a set is multiplied by a constant number, then the mean of the modified data is equal to the initial mean _____ and the median of the modified data is equal to the initial median _____

   _____.

   c. If the largest value in a data set is replaced by a smaller value, the mean of the modified data is _____

   _____

   and the median of the modified data set is _____

   _____.

3. Run MEAN-MEDIAN and enter the data set: 2, 4, 6, 9, 12, 15.

   a. Modify the data by adding +5 to each value. Record the results.

   INITIAL DATA _____

   MEAN _____ MEDIAN _____

   MODIFIED DATA (+5)_____

   MEAN _____ MEDIAN _____

   b. Run MEAN-MEDIAN with the data set: 7, 9, 11, 14, 17, 20.

   Modify the data by subtracting –5 from each data value. Record the results.

   INITIAL DATA _____

   MEAN _____ MEDIAN _____

   MODIFIED DATA (−5)_____

   MEAN _____ MEDIAN _____

   c. Show how to obtain the mean of the initial data set from the mean of the modified data set.

   _____

   _____

   d. Show how to obtain the median of the initial data set from the median of the modified data set.

   _____

   _____

4. Run MEAN-MEDIAN using the data set: 2, 4, 6, 9, 12, 15.

   a. Modify the data by multiplying each value by +4. Record the results.

   INITIAL DATA _____

   MEAN_____MEDIAN_____

   MODIFIED DATA (4×)_____

   MEAN_____MEDIAN_____

   b. Run MEAN-MEDIAN using the data set: 8, 16, 24, 36, 48, 60. Modify the data set by multiplying each value by 0.25 (dividing by 4). Record the results.

   INITIAL DATA _____

   MEAN _____ MEDIAN _____

   MODIFIED DATA (0.25×) _____

MEAN _____ MEDIAN _____

c. Show how to obtain the mean of the initial data set from the mean of the modified data set. _____

_____

5. The average and median daily temperature for six days in a northern city was 76° F. Daily temperature readings can be changed to a Celsius scale by the formula: C = (F − 32)(5/9); that is, a value of −32 must be added to each data value, and then the results must be multiplied by 5/9.

   a. If the six daily temperatures were 73°, 78°, 81°, 74°, 71°, and 79° on the Fahrenheit scale, use MEAN-MEDIAN program to change to the Celsius scale. (Don't forget to order the data as you enter it, add a –32, and then multiply the values using a decimal format: 5/9 = 0.5556).
   b. What is the average temperature of the city in Celsius? _____
   c. What is the median temperature in Celsius? _____

Dispersion

Dispersion refers to how spread out scores are around the mean. The sample **range** is the difference between the largest and smallest data value. The sample **variance** is the average squared deviation of the data values from their sample mean. The sample **standard deviation** is the square root of the sample variance. The formula for the sample variance is:

$$\text{Variance} = \frac{\sum(\text{data value} - \text{mean data value})^2}{\text{number of data values}} = \frac{SS}{n}$$

The numerator (top of equation) indicates that the mean of all the data values is subtracted from each data value, squared, and summed. The summing of the squared values is denoted by the symbol, Σ. This is referred to as the sum of squared deviations from the mean or simply **sum of squared deviations** (SS). The sum of squared deviations (SS) divided by the number of data values is referred to as the variance.

The standard deviation is the square root of the variance. The formula for the standard deviation is:

$$\text{Standard Deviation} = \sqrt{\frac{\text{SS}}{\text{n}}}$$

The standard deviation provides a measure in standard units of how far the data values fall from the sample mean. For example, in a **normal distribution**, 68% of the data values fall approximately one standard deviation (1 SD) on either side of the mean, 95% of the data values fall approximately two standard deviations (2 SD)

on either side of the mean, and 99% of the data values fall approximately three standard deviations (3 SD) on either side of the mean.

Basically, we should find that if a constant is added to all the values in a data set, the variance, standard deviation, and range are unchanged. If all of the values in a data set are multiplied by a constant: The variance of the modified data set is the variance of the initial data set times the constant squared. The standard deviation of the modified data set is the standard deviation of the initial data set times the constant. The range of the modified data set is the range of the initial data set times the constant. If the last value in a data set is replaced by a smaller value, then the variance, standard deviation, and the range are all decreased. If the last value in a data set is replaced by a larger value, then the variance, standard deviation, and the range are all increased. If the standard deviation of a data set is divided into each of the data values, then the standard deviation of the modified data set is 1.

The variance of the initial data set can be obtained from the variance of the modified data set by dividing the variance by the constant squared that was used as the multiplier. The standard deviation and range of the initial data set can be obtained from the standard deviation and range of the modified data set by dividing them by the constant that was used as the multiplier.

DISPERSION R Program

In the DISPERSION R program an initial data set is entered and modified in one of three ways: (1) adding a constant to each data value; (2) multiplying each data value by a constant; or (3) replacing the last value by a different number. The purpose is to observe the effect of these modifications on three measures of dispersion: range, variance, and standard deviation. The **sd** function returns the standard deviation of the vector of data values, the **var** function computes the variance, and the **range** function gives the minimum and maximum data value.

DISPERSION Program Output

```
Initial Data Set
 2 4 5 9 15 19
 Standard Deviation = 6.723095 Variance=45.2 Range=2 19

Added 10 to the Initial data
 Added Value Data
 Standard Deviation = 6.723095 Variance=45.2 Range=12 29

Multiplied 5 to the Initial data
 Multiplied Value Data
 Standard Deviation = 33.61547 Variance=1130 Range=10 95
Replaced Largest Value 19 in the Initial data with 1
```

```
Replaced Value Data
Standard Deviation = 5.215362 Variance=27.2 Range=1 15
```

DISPERSION Exercises

1. Run the DISPERSION program with the data set 2, 4, 5, 9, 15, 19. Use the data modifications below and record the results.

INITIAL DATA	S.D.	VARIANCE	RANGE
ADD			
3	_____	_____	_____
−2	_____	_____	_____
10	_____	_____	_____
5	_____	_____	_____
MULTIPLY BY			
2	_____	_____	_____
−10	_____	_____	_____
0.5	_____	_____	_____
REPLACE LAST VALUE WITH			
1	_____	_____	_____
5	_____	_____	_____
10	_____	_____	_____

2. Complete the following statements:

   a. If a constant is added to each data value, then the variance of the modified data is equal to _____, the standard deviation of the modified data set is equal to _____, and the range of the modified data set is equal to _____.
   b. If each data value is multiplied by a constant, then the variance of the modified data is equal to the initial variance _____, the standard deviation of the modified data is equal to the initial standard deviation _____, and the range of the modified data set is equal to the initial range _____.
   c. If the last value in a data set is replaced by a smaller value, the variance of the modified data is _____, the standard deviation of the modified data is _____, and the range of the modified data set is _____.

3. Run DISPERSION program again using your initial data set and modify it by dividing each data value by the standard deviation (multiply by the reciprocal). What is the standard deviation of the data? _____.

   Try it again with a new initial data set. Explain why this happens: _____
   _____
   _____
   _____

4. Run DISPERSION program with the data set 2, 4, 6, 9, 12, 15.

   a. Modify the data by adding +5 to each value. Record the results.

      INITIAL DATA _____

      S.D. _____VARIANCE _____ RANGE_____

      MODIFIED DATA_____

      S.D. _____VARIANCE_____ RANGE_____

   b. Run MODIFICATION and enter the data set 7, 9, 11, 14, 17, 20.

      Modify the data by adding –5 to each value. Record the results.

      INITIAL DATA _____

      S .D _____VARIANCE_____ RANGE_____

      MODIFIED DATA _____

      S.D._____VARIANCE_____ RANGE_____

   c. Show how to obtain the variance of the modified data set from the variance of the initial data set.

      _____

   d. Show how to obtain the standard deviation of the initial data set from the standard deviation of the modified data set.

      _____

   e. Show how to obtain the range of the initial data set values from the range of the modified data set values.

      _____

5. The variance of the daily temperatures for six days in a northern city was 14°F, the standard deviation was 3.8°F, and the range was 10° F. The daily temperature readings can be changed to a Celsius scale by using the formula $C=(F-32)(5/9)$; that is, –32 must be added to each data value and the results multiplied by 5/9 (0.5556).

   a. What is the standard deviation of the temperature in degrees Celsius? _____
   b. What is the variance of the temperature for this city in Celsius? _____
   c. What is the range of the temperature in degrees Celsius? _____

6. If the six daily temperatures were 73°, 78°, 81°, 74°, 71°, and 79° on the Fahrenheit scale, use the DISPERSION program to change them to the Celsius scale. (Don't forget to order the data as you enter it and express values in decimal form: $5/9=0.5556$).

a. What is the standard deviation of the temperature in degrees Celsius? _____
b. What is the variance of the temperature for this city in Celsius? _____
c. What is the range of the temperature in degrees Celsius? _____

Sample Size Effects

The sample **range** is the difference between the highest and lowest score in a data distribution. The sample **variance** is the square of the sample **standard deviation**. When the size of the sample increases, the range of data values will generally increase. The standard deviation with increasing sample sizes should divide the frequency distribution of data into six sections. You should be able to observe the effect of sample size on these measures of data dispersion when completing the chapter exercises.

As the sample size increases, the sample range usually increases because observations are chosen from the extreme data values in a population. As observations are added to a sample, the range of the sample cannot decrease. As observations are added to a sample, the standard deviation fluctuates in an unpredictable manner. A rough approximation of the standard deviation is the range divided by four; however, for a uniform population, this will produce an underestimate. Range divided by six better approximates the standard deviation of the normal distribution.

SAMPLE R Program

The SAMPLE R programs will create a uniform population of integers from 1 to 1,000 based on sampling without replacement [Sampling with replacement assumes that data points are returned to the population from which they were drawn. Sampling without replacement assumes that data points are *not* returned to the population from which they were drawn.] The probability of selection is affected depending upon which sampling technique is used. Various sample sizes will need to be listed in the *Samplesizes* vector. Random sampling will be repeated with new observations for each sample size listed in the vector. A summary table will be printed with the results to allow you to draw conclusions about the effect of sample size on the range and standard deviation. The ratio of the range to the standard deviation will be printed so you can look for a relationship between these two measures of dispersion.

The program can be repeated as many times as needed since the sampling is random. Each time the program is run, the results, however, will be different. This will allow you to further test your understanding and conclusions about how various sample sizes affect the range and standard deviation. The final chapter exercise computes the error one would make when using a sample estimate of the standard deviation as the population value. This exercise is used to answer the question, "Does the sample standard deviation become a more accurate estimate of the population standard deviation as the sample size increases?"

Once the matrix has been defined, the main processing loop begins. An iteration counter is used as the basis of the **for** loop to facilitate placement of values within the matrix. The first line of code within the loop creates the first sample of random integers between 1 and 1,000 from a uniform distribution. The next three lines fill the columns of the matrix at the present row (*i*) with the range, standard deviation, and range divided by the standard deviation, respectively. The matrix notation [*i*,1] represents the *i*th row and the first column. If you want to replace an entire row of values, type *outputMatrix*[*i*,]<-, followed by the assignment of the vector. Leaving a dimension blank means that you are allowing all values along that dimension to be filled with values, if enough are present in the vector being assigned to it. The content of the matrix is printed after the end of the loop.

Sample Program Output

```
   N  Range  Standard Dev.  Range/SD
  10    893         327.83      2.72
  50    937         286.18      3.27
 100    979         319.72      3.06
 200    991         286.30      3.46
 500    997         281.51      3.54
1000    997         288.87      3.45
```

SAMPLE Exercises

1. Enter the following string of sample sizes in the *Samplesize* vector and run the SAMPLE program. Record the results below.

	RANGE	STANDARD DEV.	RANGE/SD
20	_____	_____	_____
40	_____	_____	_____
60	_____	_____	_____
80	_____	_____	_____
100	_____	_____	_____
120	_____	_____	_____
140	_____	_____	_____
160	_____	_____	_____
180	_____	_____	_____
200	_____	_____	_____
220	_____	_____	_____
240	_____	_____	_____
260	_____	_____	_____
280	_____	_____	_____
300	_____	_____	_____

2. Provide short answers to the following questions.

   a. As observations are added to the sample, what happens to the sample range?

   _____

   b. What accounts for the relationship between the sample size and the sample range?

   _____

   c. Why is the sample range a less than perfect measure of the spread of the population?

   _____

   d. As observations are added to the sample, what is the relationship between the sample size and the sample standard deviation? _____

   _____

3. Run the SAMPLE program again with the same sample sizes in the *Samplesize* vector. Record the results below.

	RANGE	STANDARD DEV.	RANGE/SD
20	_____	_____	_____
40	_____	_____	_____
60	_____	_____	_____
80	_____	_____	_____
100	_____	_____	_____
120	_____	_____	_____
140	_____	_____	_____
160	_____	_____	_____
180	_____	_____	_____
200	_____	_____	_____
220	_____	_____	_____
240	_____	_____	_____
260	_____	_____	_____
280	_____	_____	_____
300	_____	_____	_____

   a. Did anything different happen the second time? _____
   If so, what was different? _____

   _____

   b. What is your final conclusion about the relationship between sample size and sample standard deviation? _____

4. The last column above indicates the standard deviations for samples of data from a uniform distribution. The range of scores divided by 4 is a rough estimate of the standard deviation of a uniform distribution. Are the standard deviations less than 4 (underestimated) as expected? Yes ____ No ___
5. What is a good estimate for the standard deviation of a normal population? _____
6. If the sample standard deviation is used as an estimate of the population standard deviation, compute the error of the estimate for each sample size in Exercise 1.

Note: ERROR = ESTIMATE − 288.67. The population standard deviation was 288.67.

Record the error of estimate for each sample size with its +/− signs in the following table.

SAMPLE SIZE	ERROR
20	_____
40	_____
60	_____
80	_____
100	_____
120	_____
140	_____
160	_____
180	_____
200	_____
220	_____
240	_____
260	_____
280	_____
300	_____

a. Does the sample standard deviation become a more accurate estimate of the population standard deviation as the sample size increases?
_____

Tchebysheff Inequality Theorem

The sample **standard deviation** is a measure of the dispersion of the sample data around the sample mean. A small standard deviation indicates less dispersion of sample data. A larger standard deviation indicates more dispersion of sample data. This understanding is also true for the **range**, which is the difference between the largest and smallest data value. However, the standard deviation provides more information about the data than the range. The standard deviation permits the formation of intervals that indicate the proportion of the data within those intervals. For example, 68 % of the data fall within +/− one standard deviation from the mean, 95 % of the data fall within +/− two standard deviations of the mean, and 99 % fall within +/− three standard deviations of the mean, in a **normal distribution**. If 100 students took a mathematics test with a mean of 75 and a standard deviation of 5, then 68 % of the scores would fall between a score of 70 and 80, assuming a normal distribution. In contrast, given the highest and lowest test scores, 90 and 50 respectively, the range of 40 only indicates that there is a 40-point difference between the highest and lowest test score, i.e., 90 − 50 = 40.

We generally assume our data is normally distributed; however, in some cases, the data distribution takes on a different shape. When this occurs, the **Tchebysheff Inequality Theorem** is helpful in determining the percentage of data between the intervals. For example, if the mean mathematics test score was 85, and the standard deviation was 5, then the Tchebysheff Inequality Theorem could be used to make a statement about the proportion of test scores that fall in various intervals around the mean, e.g., between the score interval 75 and 95, *regardless of the shape of the distribution*.

The **Tchebysheff Inequality Theorem** was developed by a Russian mathematician as a proof that given a number k, greater than or equal to 1, and a set of n data points, at least $(1 - 1/k^2)$ of the measurements will lie within k standard deviations of their mean. Tchebysheff's theorem applies to *any* distribution of scores and could refer to either sample data or the population. To apply the Tchebysheff Inequality Theorem using a population distribution, an interval is constructed which measures $k\sigma$ on either side of the mean, μ. When $k = 1$, however, $1 - 1/(1)^2 = 0$, which indicates that 0 % of the data points lie in the constructed interval, $\mu - \sigma$ to $\mu + \sigma$, which is not helpful nor useful in explaining data dispersion. However, for values of k greater than 1, the theorem appears to be informative:

k	$1 - 1/k^2$ (Percent)	Interval
1	0 (0 %)	$\mu +/- 1\sigma$
2	3/4 (75 %)	$\mu +/- 2\sigma$
3	8/9 (89 %)	$\mu +/- 3\sigma$

An example will help to better illustrate the fraction of n data points that lie in a constructed interval using the Tchebysheff theorem. Given a set of test scores with a mean of 80 and a standard deviation of 5, the Tchebysheff theorem would indicate a constructed interval with lower and upper score limits computed as follows:

Lower limit = mean − k * standard deviation
Upper limit = mean + k * standard deviation

For $k = 1$, the lower limit would be $80 - 1*5 = 75$ and the upper limit would be $80 + 1*5 = 85$. Obviously, for $k = 1$, no data points are implied between the score interval, 75–85, which makes no sense. For $k = 2$, the lower limit would be 70 and the upper limit would be 90. The Tchebysheff Inequality Theorem implies that *at least* $[1 - 1/k^2]$ of the data values are within the score interval. Thus, for the constructed interval $k = 2$, 70–90, at least $[1 - 1/(2)^2] = 1 - 1/4 = 1 - 0.25 =$ *at least* 75 % of the data points are between 70 and 90, *regardless of the shape of the data distribution*. The Tchebysheff Inequality Theorem is very conservative, applying to *any* distribution of scores, and in most situations the number of data points exceeds that implied by $1 - 1/k^2$.

The Tchebysheff Inequality Theorem is generally applied to populations and intervals formed around the population mean using k population standard deviations, where k ranges from 1 to 4. In practice, however, one rarely knows the population parameters (population means and standard deviations). In some instances, the population parameters are known or at least can be estimated. For example,

a nationally normed test booklet would contain the population mean and standard deviation, typically called "test norms." Researchers often use tests to measure traits and characteristics of subjects and publish test sample means and standard deviations. In this instance, an average of the results from several published studies would yield a reasonable estimate of the population parameters. In finite populations where every observation is known and recorded, the population parameters are readily obtainable using computer statistical packages to analyze the data. In a few instances, dividing the range of scores by six provides a reasonable estimate of the population standard deviation as an indicator of data dispersion.

Since the sample mean and standard deviation are estimates of the population parameters, the Tchebysheff Inequality Theorem can be used with sample data. We therefore can test whether or not the Tchebysheff Inequality Theorem is useful for describing data dispersion and compare it to the normal distribution percentages where approximately $1\sigma = 68\ \%$, $2\sigma = 95\ \%$, and $3\sigma = 99\ \%$. In the TCHEBYSHEFF program, samples will be selected from four different populations: uniform, normal, exponential, or bimodal. The four different distributions are functions within the R program; however, they can be created in other programming software by computing a value for X(i), which is a data vector, using the following equations and functions (RND = round a number; COS = cosine of a number; SQR = square root of a number; LOG = logarithm of a number):

```
Uniform:      X(i)  =  1 + 9*RND
Normal:       X(i)  =  COS (6.2832*RND) * SQR(−2*LOG (RND))
Exponential:  X(i)  =  −LOG (RND)
Bimodal:      X(i)  =  (2 + SQR (4 – (8* (1 – RND))))/2
              If RND ≥ 0.5
              X(i)  =  (2 − SQR (4 – (8*RND)))/2
```

The Tchebysheff Inequality Theorem provides a lower bound for the proportion of sample data within intervals around the mean of any distribution. The Tchebysheff Inequality Theorem is true for all samples regardless of the shape of the population distribution from which they were drawn. The Tchebysheff lower bound is often a conservative estimate of the true proportion in the population. We would use the standard deviation to obtain information about the proportion of the sample data that are within certain intervals of a normally distributed population. The population mean and standard deviation are estimated more accurately from large samples than from small samples.

TCHEBYSHEFF R Program

The program will require specifying the sample size and distType. The TCHEBYSHEFF program will select a random sample, compute the sample mean and standard deviation, and determine the percentage of the observations within 1.5, 2, 2.5, and 3 standard deviations of the mean (Kvals). The lower bound for the

percentage of data within the interval given by the Tchebysheff Inequality Theorem will also be printed to check whether the theorem is true for the sample data.

The program takes random samples from different shaped distributions. The "Uniform" selection chooses a random sample of *SampleSize* from a uniform distribution that falls between the values of 1 and 10. "Normal" creates a random sample from a normal distribution with a mean of 0 and standard deviation of 1. "Exponential" creates a random sample from an exponential distribution with a mean of 1. Finally, "Bimodal" creates a random sample from a bimodal distribution (made up of an equal number of points chosen from two adjacent normal distributions). Whichever distribution type is input, the standard deviation and mean are obtained, and then the matrix is filled with values representing the Tchebysheff intervals, the percent of observations in the sample falling within the interval, and the value of the Tchebysheff Lower Bound.

TCHEBYSHEFF Program Output

```
Uniform N=50 Sample Mean 5.17 Sample Std Dev 2.51

  K       Interval     % Obs    Tcheby
 1.5   1.4 to 8.9        86        56
 2.0   0.16 to 10       100        75
 2.5   -1.1 to 11       100        84
 3.0   -2.4 to 13       100        89

Normal N=50 Sample Mean 0.04 Sample Std Dev 0.97

  K         Interval    % Obs    Tcheby
 1.5    -1.4 to 1.5       88        56
 2.0      -1.9 to 2       98        75
 2.5    -2.4 to 2.5       98        84
 3.0      -2.9 to 3      100        89

Exponential N=50 Sample Mean 1.15 Sample Std Dev 1.14

  K           Interval    % Obs    Tcheby
 1.5     -0.56 to 2.9       90        56
 2.0      -1.1 to 3.4       92        75
 2.5        -1.7 to 4       96        84
 3.0      -2.3 to 4.6      100        89

Bimodal N=50 Sample Mean 0.92 Sample Std Dev 0.62

  K           Interval    % Obs    Tcheby
 1.5         0 to 1.8       88        56
 2.0     -0.31 to 2.1       96        75
 2.5     -0.62 to 2.5       98        84
 3.0     -0.93 to 2.8      100        89
```

TCHEBYSHEFF Exercises

1. Run TCHEBYSHEFF for the sample size =50 and distType = "Uniform" specified in the program. Then replace the distType for the other distribution types.

UNIFORM N = 50 Sample Mean _____ Sample St. Dev. _____

K	INTERVAL	% OBS. IN INT.	TCHEBY. LOWER BOUND
1.5	_____	_____	_____
2.0	_____	_____	_____
2.5	_____	_____	_____
3.0	_____	_____	_____

NORMAL N = 50 Sample Mean _____ Sample St. Dev. _____

K	INTERVAL	% OBS. IN INT.	TCHEBY. LOWER BOUND
1.5	_____	_____	_____
2.0	_____	_____	_____
2.5	_____	_____	_____
3.0	_____	_____	_____

EXPONENTIAL N = 50 Sample Mean _____ Sample St. Dev. _____

K	INTERVAL	% OBS. IN INT.	TCHEBY. LOWER BOUND
1.5	_____	_____	_____
2.0	_____	_____	_____
2.5	_____	_____	_____
3.0	_____	_____	_____

BIMODAL N = 50 Sample Mean _____ Sample St. Dev. _____

K	INTERVAL	% OBS. IN INT.	TCHEBY. LOWER BOUND
1.5	_____	_____	_____
2.0	_____	_____	_____
2.5	_____	_____	_____
3.0	_____	_____	_____

2. Are the Tchebysheff lower bound values always correct in the table? _____
_____

3. The Tchebysheff lower bound is very conservative; it is often a lower bound far below the actual percentage of data in the interval. For which population is the Tchebysheff lower bound the least conservative? Run TCHEBYSHEFF several times with different sample sizes to verify your conclusion.
_____

_____

4. The actual population means and standard deviations of the four populations are:

	POP. MEAN	POP. ST. DEV.
UNIFORM	5.5	2.60
NORMAL	0	1
EXPONENTIAL	1	1
BIMODAL	1	0.707

The sample means and standard deviations computed in TCHEBYSHEFF can be used as estimates of these population parameter values.

a. Run TCHEBYSHEFF to complete the following table. Recall that

ERROR = SAMPLE ESTIMATE – POPULATION VALUE.

SAMPLE SIZE	SAMPLE MEAN	POP. MEAN	ERROR	SAMPLE ST. DEV.	POP ST. DEV.	ERROR
UNIFORM						
20	_____	_____	_____	_____	_____	_____
50	_____	_____	_____	_____	_____	_____
100	_____	_____	_____	_____	_____	_____
NORMAL						
20	_____	_____	_____	_____	_____	_____
50	_____	_____	_____	_____	_____	_____
100	_____	_____	_____	_____	_____	_____
EXPONENTIAL						
20	_____	_____	_____	_____	_____	_____
50	_____	_____	_____	_____	_____	_____
100	_____	_____	_____	_____	_____	_____
BIMODAL						
20	_____	_____	_____	_____	_____	_____
50	_____	_____	_____	_____	_____	_____
100	_____	_____	_____	_____	_____	_____

b. Is there a relationship between the sample size and the absolute value of the error in the estimates of the mean? _____

_____

c. Is there a relationship between the sample size and the absolute value of the error in the estimates of the standard deviation? _____

_____

Normal Distribution

In a normal population, referred to as normal bell-shaped curve, the proportion of data within intervals around the mean is known. The proportion of sample data within k standard deviations around the mean for $k = 1$, 2, and 3 are as follows:

INTERVAL		DATA PERCENT
LOWER LIMIT	UPPER LIMIT	IN THE INTERVAL
$\mu - 1\sigma$	$\mu + 1\sigma$	68%
$\mu - 2\sigma$	$\mu + 2\sigma$	95%
$\mu - 3\sigma$	$\mu + 3\sigma$	99%

If a large random sample is selected from a normal population, these lower and upper intervals will approximate the percent of data within the intervals of a normal bell-shaped curve. The proportions of sample data should be good approximations if the population is approximately normal, symmetrical, and unimodal. If the population is non-normal, the bell-shaped curve may not provide results close to these proportions of sample data.

The different population distribution functions in R permit the random sampling of data from different population distributions. The different population values can be adjusted and the population parameters for the different distributions are:

	MEAN	STANDARD DEVIATION
UNIFORM	5.5	2.60
NORMAL	0	1
EXPONENTIAL	1	1
BIMODAL	1	0.707

The normal distribution gives an approximation of the proportion of a sample that is within one, two, and three standard deviations of the population mean if the sample is large and is chosen at random from a normal population. The percentages of data within one, two, and three standard deviations of the mean for a normal population are respectively 68%, 95%, and 99%. The program does not give good approximations for samples chosen from uniform, exponential, or bimodal populations nor good approximations for small samples.

BELL-SHAPED CURVE R Program

The BELL-SHAPED CURVE program will create large sample data sets chosen at random from the four population types: uniform, normal, exponential, and bimodal. The program will determine the proportion of the sample within the specified intervals. The tabled output should allow you to check the accuracy of the percentages for large samples from the population distribution types, especially a normal population. The proportions within the specified intervals for smaller samples can also be checked.

The BELL-SHAPED CURVE program is similar to the TCHEBYSHEFF program except **Kvals** is set to intervals using percentages which are specified as *Kvals <− c(68,95,99)*; which corresponds to 1, 2, or 3 standard deviations from the population mean, respectively. The sample size is initially set at *SampleSize <− 50*.

BELL-SHAPED CURVE Program Output

```
Uniform N=50 Sample Mean=5.38 Sample Std Dev=2.87

        Interval      % Observed    % Predicted
  K=1    2.5 to 8.2     54           68
  K=2   -0.37 to 11    100           95
  K=3   -3.2 to 14100   99

  Normal N=50 Sample Mean=-0.16 Sample Std Dev=0.98

        Interval      % Observed    % Predicted
  K=1   -1.1 to 0.83    60           68
  K=2   -2.1 to 1.8     98           95
  K=3   -3.1 to 2.8    100           99

  Exponential N=50 Sample Mean=1.01 Sample Std Dev=1.17

        Interval      % Observed    % Predicted
  K=1   -0.16 to 2.2   88           68
  K=2   -1.3 to  3.3   92           95
  K=3   -2.5 to  4.5   98           99

  Bimodal N=50 Sample Mean=1.05 Sample Std Dev=0.69

        Interval      % Observed    % Predicted
  K=1   0.36 to 1.8     60           68
  K=2  -0.33 to 2.4     98           95
  K=3     -1 to 3.1    100           99
```

Normal Distribution Exercises

1. Run BELL-SHAPED CURVE for a sample size of 250, and complete the following tables. Note: ERROR = % OBSERVED – % PREDICTED.

   a. Compute the lower and the upper limits of the interval for the uniform population with $k = 3$, using the following formula:

      LOWER LIMIT = MEAN – K*STANDARD DEVIATION = _____

      UPPER LIMIT = MEAN + K*STANDARD DEVIATION = _____

   b. Which population produced a sample that is approximated best by the Normal Bell-Shaped Curve? _____

   c. Run BELL-SHAPED CURVE program again for sample size 500 to test your conclusions. Comment on the results.

      _____

UNIFORM: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____
NORMAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____
EXPONENTIAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____
BIMODAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

2. Run BELL-SHAPED CURVE program again for a sample size of 1,000 and complete the tables.

UNIFORM: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____
NORMAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____
EXPONENTIAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____
BIMODAL: N =		Sample Mean =	Sample Std Dev =	
k	INTERVAL	% OBSERVED	% PREDICTED	ERROR
1	_____	_____	68	_____
2	_____	_____	95	_____
3	_____	_____	99	_____

a. Compare the results for the samples of size 1,000 with the results for samples of size 250. For which sample size (250 or 1,000) were the approximations best for the normal population? _____
b. Was there more error in the nonnormal populations (uniform, exponential, or bimodal)? Yes _____ No _____

3. The uniform, exponential, and the bimodal populations all have shapes that are very different from the normal population.

a. In what way is the uniform population different from the normal population?

_____

b. In what way is the exponential population different from the normal population?

_____

c. In what way is the bimodal population different from the normal population?

_____

d. Which population type had the most error?

_____

4. Run BELL-SHAPED CURVE program for a sample size of n = 20 for all four populations.

a. Enter the percent of data for each value of *k* in the table.

k	UNIFORM	NORMAL	EXPONENTIAL	BIMODAL
1	_____	_____	_____	_____
2	_____	_____	_____	_____
3	_____	_____	_____	_____

b. Comment on the accuracy of the Normal Bell-Shaped Curve for small samples.
Hint: Is the Normal Bell-Shaped Curve still better than the others?

_____

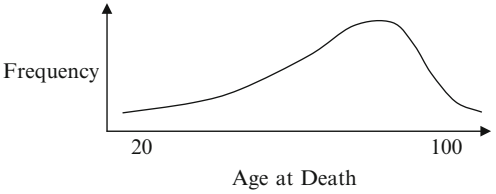
Central Limit Theorem

In some instances, the normal distribution may not be the type of distribution we obtain from sample data when studying research variables and/or the population data may not be normally distributed on which we base our statistics. The normal

probability distribution however is still useful because of the Central Limit Theorem. The Central Limit Theorem states that as sample size increases a sampling distribution of a statistic will become normally distributed even if the population data is not normally distributed. The sampling distribution of the mean of any nonnormal population is approximately normal, given the Central Limit Theorem, but a larger sample size might be needed depending upon the extent to which the population deviates from normality.

Typically, a smaller sample size can be randomly drawn from a *homogeneous* population, whereas a larger sample size needs to be randomly drawn from a *heterogeneous* population, to obtain an unbiased sample estimate of the population parameter. If the population data are normally distributed, then the sampling distribution of the mean is normally distributed; otherwise larger samples of size N are required to approximate a normal sampling distribution. The sampling distribution of the mean is a probability distribution created by the frequency distribution of sample means drawn from a population. The sampling distribution, as a frequency distribution, is used to study the relationship between sample statistics and corresponding population parameters.

The Central Limit Theorem is useful in statistics because it proves that sampling distributions will be normally distributed regardless of the shape of the population from which the random sample was drawn. For example, a physician is studying the life expectancy of adults after being diagnosed with cancer. She is going to do a statistical analysis on the data concerning age at death and needs a theoretical probability distribution to model the adult ages. Since most of the adults lived to an advanced age due to new cancer treatments, she realizes that the population of ages is skewed to the left (see Figure below).



The physician doesn't know whether a mathematical function would best describe this population distribution, but would like to test mean differences in age at death between normal adults and adults with cancer. Fortunately, she can use a sampling distribution of sample means, which doesn't require exact knowledge of the population distribution to test her hypothesis. The Central Limit Theorem, which is based on the sampling distribution of statistics, permits the use of the normal distribution for conducting statistical tests.

Sampling distributions of the mean from non-normal populations approach a normal distribution as the sample size increases, which is the definition of the Central Limit Theorem. The frequency distribution of sample mean based on samples of size N randomly drawn from a population is called the sampling distribution of the mean. The sampling distribution of the mean is a normally distributed

probability distribution. The mean of the sampling distribution of the mean is equal to the mean of the population being sampled. The variance of the sampling distribution of the mean is equal to the variance of the population being sampled divided by the sample size N. Sampling distributions of the mean from normal populations are normally distributed.

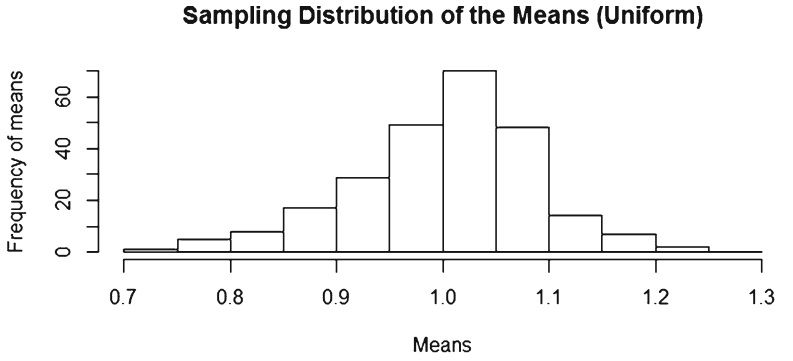
CENTRAL R Program

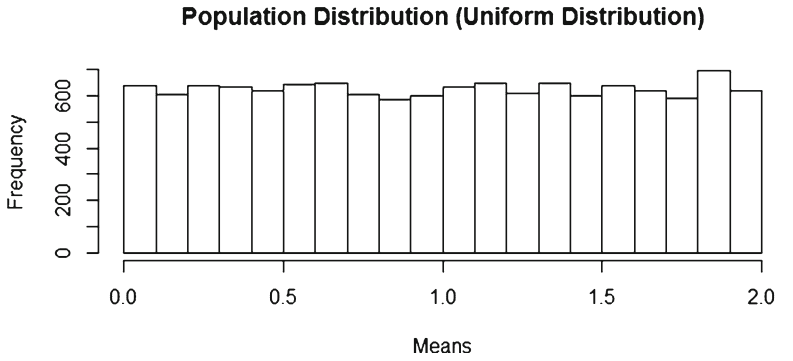
The CENTRAL program will graph the sampling distribution of the mean for samples of a given size N. The random samples will be taken from one of four different population types: uniform, normal, exponential, or bimodal. The frequency distribution of the sample means can be based on an infinite number of samples, but the initial value in the program is set at 250 samples. The sampling distributions of the mean approaches a normal distribution, as sample size increases. Because of this, you will be able to observe how the underlying population distribution type does not affect the normality of the sampling distribution of the mean. The program output shows that the sampling distribution is normally distributed even when data comes from many different types of population distributions.

The program begins by initializing the user-defined variables, including selection of the underlying distribution type from which samples are drawn. The main loop iterates for the number of desired replications, creating a sample of the appropriate size from the appropriate distribution. The parameters for the samples are set so that most of the sampling distributions of the mean should fall between 0 and 2 for ease of comparison. After each sample is selected, the mean is calculated and added to the vector of sample means. The entire sample is added to a vector that contains all the raw data from every sample, thereby creating a very large, single sample. When replications are finished, an output vector is created to display the mean and variance for the population distribution and sampling distribution of the mean. Two histograms are graphed, one for the sampling distribution and one for the population distribution.

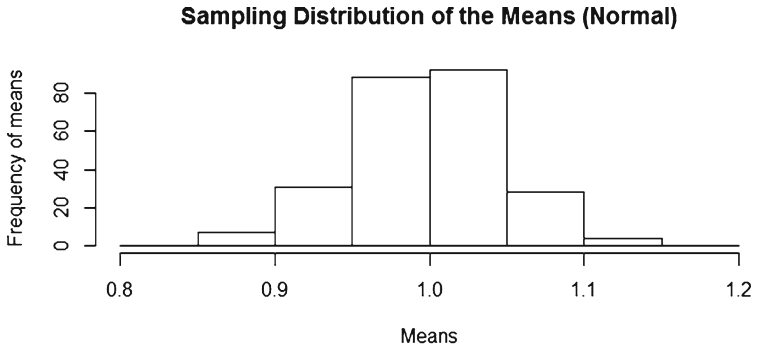
CENTRAL Program Output

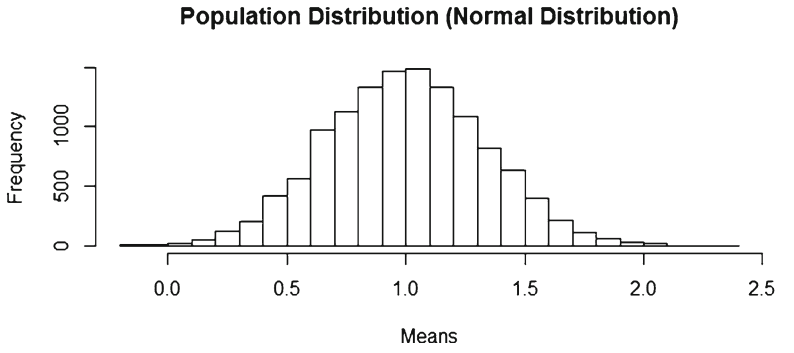
```
                        Inputvalues
Sample Size                      50
Number Replications             250
Distribution Type           Uniform

Sampling Distribution Mean=1.00107 Variance=0.00731183
Uniform Distribution Mean=0.9468208 Variance 0.3919674
```





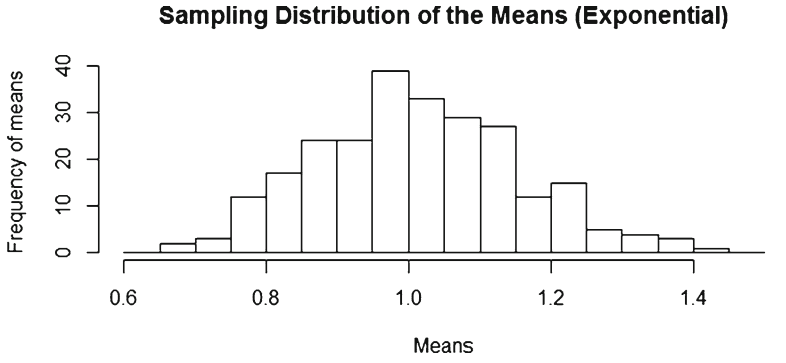

```
                        Inputvalues
Sample Size                      50
Number Replications             250
Distribution Type           Normal

Sampling Distribution Mean=0.9975896 Variance=0.002228887
Normal Distribution Mean=1.010424 Variance 0.1147131
```

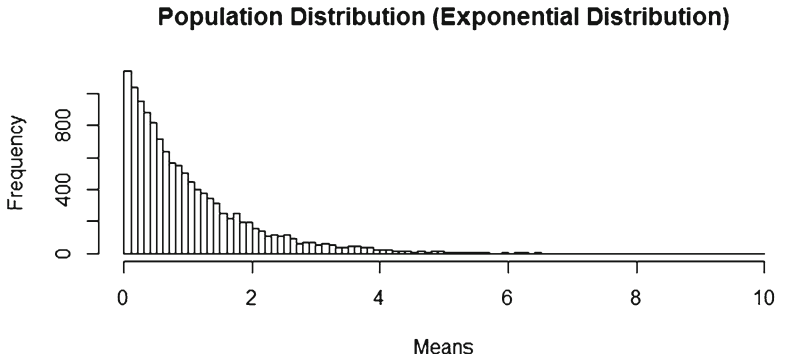




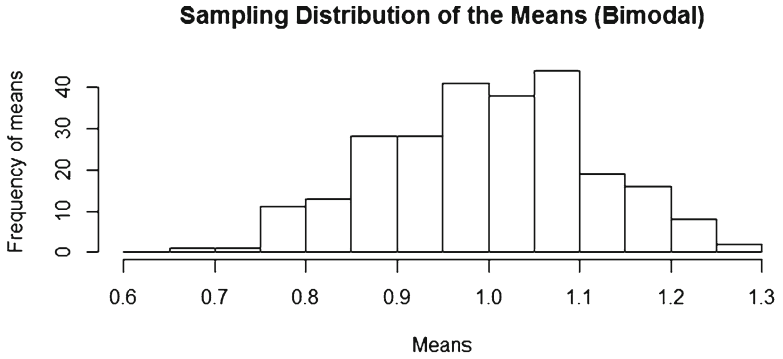
	Inputvalues
Sample Size	50
Number Replications	250
Distribution Type	Exponential

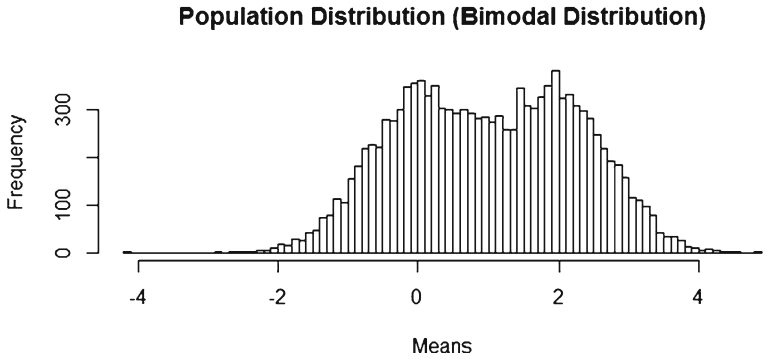
Sampling Distribution Mean=1.013069 Variance=0.02057427
Exponential Distribution Mean=1.132056 Variance 0.8740562






```
                        Inputvalues
Sample Size                      50
Number Replications             250
Distribution Type           Bimodal

Sampling Distribution Mean=0.9995466 Variance=0.01314928
Bimodal Distribution Mean=1.243848 Variance 1.757248
```





Central Limit Theorem Exercises

1. Run CENTRAL for sample size N = 5 for each of the four population types. Enter the tabled output of results below. Print the sampling distribution and population distribution graphs for each of the population distribution types.

UNIFORM			
	POPULATION	THEORETICAL	SAMPLING DISTRIBUTION
MEAN	_____	_____	_____
VARIANCE	_____	_____	_____
NORMAL			
	POPULATION	THEORETICAL	SAMPLING DISTRIBUTION
MEAN	_____	_____	_____
VARIANCE	_____	_____	_____

EXPONENTIAL

	POPULATION	THEORETICAL	SAMPLING DISTRIBUTION
MEAN	_____	_____	_____
VARIANCE	_____	_____	_____

BIMODAL

	POPULATION	THEORETICAL	SAMPLING DISTRIBUTION
MEAN	_____	_____	_____
VARIANCE	_____	_____	_____

a. Which of the sampling distributions are the most like a normal distribution?

_____

b. Which sampling distribution is most different from a normal distribution?

_____

c. Are the sampling distributions approximately normal regardless of the shape of the underlying population? YES _____ NO _____

2. Run the CENTRAL program for samples of size N = 30 for each population type. Draw a rough graph of each sampling distribution of the mean.

_____ UNIFORM

_____ NORMAL

_____ EXPONENTIAL

_____ BIMODAL

a. Are the sampling distributions for sample size N = 5 and N = 30 different?

YES _____ NO _____

b. Are the population means and sampling distribution means the same for samples of size N = 30? Note: ERROR = THEORETICAL − SAMPLING.

					ERROR
Uniform	YES	_____	NO	_____	_____
Normal	YES	_____	NO	_____	_____
Exponential	YES	_____	NO	_____	_____
Bimodal	YES	_____	NO	_____	_____

c. Are the population variances and sampling distribution variances the same for samples of size N = 30? Note: ERROR = THEORETICAL − SAMPLING.

					ERROR
Uniform	YES	_____	NO	_____	_____
Normal	YES	_____	NO	_____	_____
Exponential	YES	_____	NO	_____	_____
Bimodal	YES	_____	NO	_____	_____

3. Run the CENTRAL program again, but this time select a sample size of 100. Answer the following questions.

   a. Is a sample size of 100 sufficiently large to produce a sampling distribution of the mean that is approximately normal regardless of the population type?

      YES _____ NO _____

   b. Is the mean of each population type related to the mean of the sampling distribution of the mean?

      YES _____ NO _____

   c. Is the variance of each population type related to the variance of the sampling distribution of the mean?

      YES _____ NO _____

   d. Would the means and variances for a single sample be equal to the mean and variance of each population type?

      YES _____ NO _____

True or False Questions

Central Tendency

T F a. Changing the largest value in a data set to a value four times as large does not effect the median.

T F b. If the mean of a data set is subtracted from every value of the data set, the mean of the modified data is equal to the mean of the initial data set.

T F c. Multiplying each value in a data set by ½ causes the median of the modified data to be twice as large as the median of the initial data.

T F d. A data set is modified by adding a constant to each value; the mean of the initial data set can be found by subtracting the constant from the mean of the modified data set.

T F e. A data set is modified by multiplying all values by 5. The median is now 33.5. The median of the initial data set can be found by multiplying the new median by 0.2

Dispersion

T F a. Changing the last value in a data set to a value four times as large always multiplies the range by four.

T F b. Adding five to every data value does not affect the standard deviation.

T F c. Multiplying each value in a data set by 1/3 causes the variance of the modified data to be 1/6 of the original variance.

T F d. If a data set is modified by adding a constant to each value; the range of the initial data set can be found by subtracting the constant from the range of the modified data set.

T F e. If a data set is modified by multiplying all values by 5 and the standard deviation is now 33.5; the standard deviation of the original data set can be found by multiplying the new standard deviation by 0.2.

Sample Size Effects

T F a. The sample range usually decreases as the sample size increases.

T F b. The sample standard deviation decreases as the sample size increases.

T F c. A rough approximation of the range is four times the variance.

T F d. The range of a small sample is usually less than the range of a larger sample from the same population.

T F e. The standard deviation of a uniform population is underestimated if one-fourth of the range is used for the estimate.

Tchebysheff Inequality Theorem

T F a. The Tchebysheff Inequality Theorem gives the approximate percentage of observations within certain intervals of the population.

T F b. If $k = 1$, the Tchebysheff Inequality Theorem states that at least 0% of the data are within one standard deviation of the mean.

T F c. The Tchebysheff Inequality Theorem is always true and does not depend on the shape of the population.

T F d. The error in the estimate of the standard deviation is usually larger for larger samples.

T F e. The Tchebysheff Inequality Theorem states that in a sample, at least 93.75 % of the data are within four standard deviations of the mean.

Normal Distribution

T F a. The Normal Bell-Shaped Curve gives a lower limit for the percentage of the data set within certain intervals.

T F b. The Normal Bell-Shaped Curve is based on what is known about a normal population and assumes that the sample is large and unimodal.

T F c. The Normal Bell-Shaped Curve gives accurate estimates for an exponential population.

T F d. The Normal Bell-Shaped Curve gives good estimates for small data sets.

T F e. Although the Normal Bell-Shaped Curve gives a more precise statement than the TCHEBYSHEFF lower bound, the Normal Bell-Shaped Curve has the disadvantage that it does not apply to populations of all possible shapes.

Central Limit Theorem

T F a. If a population is normal, the sampling distribution of the mean will also be normal.

T F b. A frequency distribution of 100 sample means for samples of size N drawn from a population is called the sampling distribution of the mean.

T F c. The variance of sampling distribution of mean is equal to the variance of the underlying population divided by the sample size.

T F d. The mean of a sampling distribution of mean is not normally distributed if the underlying population is not normally distributed.

T F e. A sampling distribution of mean is a probability distribution.