# TI™ Technology Manual to Accompany

# Mind on Statistics

## FIFTH EDITION

**Jessica M. Utts**
University of California, Irvine
Irvine, CA

**Robert F. Heckard**
Pennsylvania State University
State College, PA

Prepared by

**Melissa M. Sovak**
California University of Pennsylvania, California, PA

CENGAGE Learning®

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

ISBN-13: 978-1-285-83862-5
ISBN-10: 1-285-83862-9

**Cengage Learning**
200 First Stamford Place, 4th Floor
Stamford, CT 06902
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at: **www.cengage.com/global**.

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, visit **www.cengage.com**.

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com**.





Printed in the United States of America
1 2 3 4 5 6 7 17 16 15 14 13

# Contents

# Chapter 1
# Introduction to the TI-83 Plus Silver Edition and the TI-84 Plus Silver Edition

## 1.1 Getting Started

This chapter represents a brief introduction to the TI-83 Plus Silver Edition (hereafter referred to the TI-83 Plus SE) and the TI-84 Plus Silver Edition (hereafter referred to the TI-84 Plus SE). Basic commands, techniques and the use of lists are discussed briefly in this introduction. Detailed descriptions of built in calculator functions are given in the TI-83 Plus SE and TI-84 Plus SE guidebooks.

After reading this chapter you should be able to:

1. Turn the calculator on and off.
2. Adjust the display contrast.
3. Evaluate an expression.
4. Use last entry to edit an expression and evaluate and expression.
5. Access menu options.
6. Display the mode settings.
7. Graph a function.
8. Enter a list.
9. Plot a statistical data set.
10. Save a list using a descriptive name.
11. Clear lists.

## 1.2 Features

The keypad on the TI-83 Plus SE and TI-84 Plus SE are virtually identical. The TI-84 Plus SE, TI-83 Plus SE, and the TI-83 Plus are keystroke-for Keystroke compatible. The keyboard is divided into zones: graphing keys, editing keys, advanced function keys, and scientific calculator keys. The graphing keys access the interactive graphing features and are located on the first row at the top of the keyboard. The editing keys allow you to edit expressions and values and are located on the second and third rows below the graphing keys. The advanced function keys display menus that access the advanced functions: MATH, APPS, PRGM, VARS and are located on the fourth row below the graphing keys. The scientific calculator

keys access the capabilities of a standard scientific calculator and are the remaining keys located on rows five through ten.

The TI-83 Plus SE and TI-84 Plus SE uses *Flash* technology, which lets you upgrade to future software versions without buying a new graphing handheld calculator. As new software becomes available, you can electronically upgrade your TI-84 Plus from the Internet. The primary differences between the TI-83 Plus SE and the TI-84 Plus SE occur in

1. The TI-83 Plus SE is preloaded with one application. The TI-84 Plus SE is preloaded with numerous applications.
2. The TI-83 Plus SE uses a TI-Graph link that is available as an accessory from TI. The TI-84 Plus SE comes with a USB unit-to-unit cable to connect and communicate with another TI-84 Plus Silver Edition. With TI Connect software and a USB computer cable, you can also link the TI-84 Plus SE to a personal computer.

## 1.3 The Basics

**Keystrokes Introduced**

1. [ON] turns on the calculator.
2. [2nd] [OFF] turns the calculator off.
3. [2nd] [▲] darkens the screen; [2nd] [▼] lightens the screen.
4. [2nd] [MEM] accesss the MEMORY menu.
5. [2nd] [QUIT] returns to the home screen.
6. [ENTER] may be used to evaluate an expression or execute a menu option.
7. [2nd] [ENTER] recalls the last entry.
8. [STAT] displays the STAT menu.
9. [ALPHA] [▼] moves the cursor down one screen at a time.
10. [MODE] displays the mode settings.
11. [Y=] displays the Y= editor.
12. [WINDOW] displays the current window variable values.
13. [GRAPH] displays the graph of a selected function.
14. [ZOOM] >ZStandard sets the standard window variables.

To turn on the calculator press the [ON] key, and the key sequence [2nd] [OFF] turns the calculator off. There is a battery saving feature on the calculator that will automatically turn off the TI-83 Plus SE and the TI-84 Plus SE.

The 2nd key located on the top left and the up and down cursor movement keys located on the top right portion of the keypad are used to adjust the screen contrast. The keystrokes 2nd ▲ darken the screen and 2nd ▼ lighten the screen. This keystroke sequence, when repeated, will continuously darken or lighten the screen.

You can adjust the display contrast to suit your viewing angle and lighting conditions. As you change the contrast setting, a number from 0 (lightest) to 9 (darkest) in the top-right corner indicates the current level. You may not be able to see the number if contrast is too light or too dark. Both the TI-83 Plus SE and the TI-84 Plus SE have 40 contrast settings, so each number 0 through 9 represents four settings. When the batteries are low, a low-battery message is displayed when you turn on the calculator.

Variables (real or complex number, list, matrix, Y= variable, program, Apps, AppVars, picture, graph database, or string) stored in the calculator may be selectively deleted. The 2nd MEM keystrokes access the MEMORY menu as shown in Figure 1.1.

Figure 1.1

## Home Screen

The home screen is the primary screen of the TI-83 Plus SE and the TI-84 Plus SE. The appearance of the cursor indicates what will happen when you press the next key or select the next menu item to be pasted as a character on the home screen. On this screen, you may enter instructions to execute and evaluate expressions. Answers are displayed on this home screen. The blinking rectangular cursor, ■, indicates the calculator is ready to accept commands. To return to the home screen from any other screen, use 2nd QUIT.

## Evaluating Expressions

The order of operations applies to all expressions entered into the calculator. Parentheses should be used to ensure the desired order of operations, with the grey negation key (-), being used for negation. The grey negation key is located on the bottom row, column four of the keyboard. After entering an expression, press the ENTER key to evaluate the expression. Figures 1.2 and 1.3 illustrate several

arithmetic calculations.

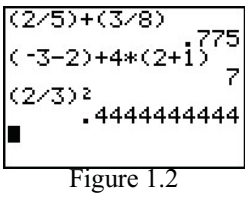


Figure 1.2

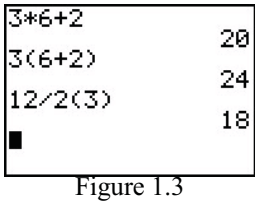


Figure 1.3

## Last Entry

When you press [ENTER] on the home screen to evaluate an expression or execute an instruction, the expression or instruction is placed in a storage area called ENTRY (last entry). When you turn off the TI-84 Plus, ENTRY is retained in memory. To recall ENTRY, press [2nd] [ENTER]. The last entry is pasted to the current cursor location, where you can edit and execute it. On the home screen or in an editor, the current line is cleared and the last entry is pasted to the line.

**Example 1.1** A dataset consists of handspan values in centimeters for six females; the values are 21, 19, 20, 20, 29, and 19. The mean is the numerical average, calculated as the sum of the data values divided by the number of values. (Utts/Heckard, Statistical Ideas and Methods, p32)

**Follow these steps** to learn the process of editing an expression.

1. Enter the data and determine the mean.

   Enter the data as shown in Figure 1.4. Press [ENTER] to evaluate the expression.

2. Edit the expression.

   An error was found in the data recording. Examination of the data indicates that the 29 should actually be a 22. Press [2nd] [ENTRY] to display the expression once again. Use the up arrow key, [▲], placing the cursor on the 9 of the value 29. Change the 29 to 22. Press [ENTER] to evaluate the revised expression. This process is illustrated in Figures 1.4, 1.5, and 1.6.

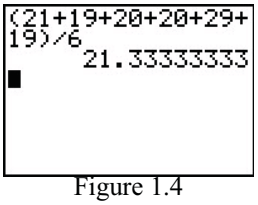


Figure 1.4

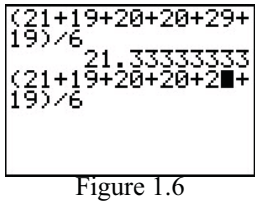


Figure 1.6

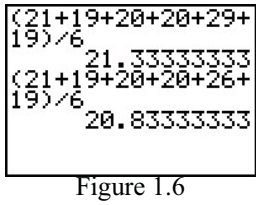


Figure 1.6

## Menus

You can access the TI-83 Plus SE and TI-84 Plus SE operations using menus. When

you press a key or key combination to display a menu, one or more menu names appear on the top line of the screen. When you press a key that displays a menu, that menu temporarily replaces the screen where you are working. For example, when you press [STAT], the STAT menu is displayed as a full screen, as shown in Figure 1.7. The current, or active, menu will be highlighted or darkened. The left and right arrow keys, [◄] and [►], move the cursor to the other menu options. To select a menu option, press the number of the menu option desired, or move the cursor up or down with the arrow keys, [▲] and [▼] to highlight the desired selection and press [ENTER]. Observe that if the left-most menu option is highlighted, pressing the left arrow, [◄], causes the cursor to highlight the right-most menu option. If more than a screen-full of menu options press [ALPHA] [▼] to move down one screen at a time.

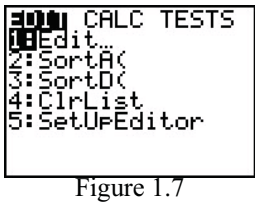


Figure 1.7

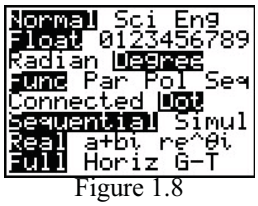


Figure 1.8

## Display Modes

Mode settings control how the TI-83 Plus SE and TI-84 Plus SE displays and interprets numbers and graphs. Mode settings are retained by the Constant Memory feature when the TI-83 Plus SE and TI-84 Plus SE is turned off. All numbers, including elements of matrices and lists, are displayed according to the current mode settings. The [MODE] key, 2nd row, 2nd column, is used to view and/or change the mode settings. To select a particular setting, move the cursor with the arrow keys to the desired option and press [ENTER] to highlight that option. Once you have selected the desired settings, press [2nd] [QUIT]. Recommended settings are shown in Figure 1.8.

## Graphing

You can store, graph, and analyze up to 10 functions, up to six parametric functions, up to six polar functions, and up to three sequences. You can use DRAW instructions to annotate graphs. Mode settings must be changed appropriately.

### Functions

You can store, graph, and analyze up to 10 functions, up to six parametric functions, up to six polar functions, and up to three sequences. You can use DRAW instructions to annotate graphs. Mode settings must be changed appropriately.

**Example 1.2** Normal random variables are the most common type of continuous

random variables. The bell-shaped normal curve illustrates the distribution of these normal random variables. (Utts/Heckard, Statistical Ideas and Methods, p268)

**Follow these steps** to graph a normal probability function $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.

1. Enter the function.

   Press [Y=], row 1, column 1, to enter the function, as shown in Figure 1.9. Press $(1/\sqrt{(2\pi)})e\hat{}(-1/2)(x^2)$. The left and right parenthese are located on row 6. Press [2nd] $\pi, \pi$ is located on the 5th row, right column above the [^] key. Press [2nd] $e, e$ is located on the 8th row, left column above the [LN] key. Be sure to use the grey negation key when you enter $(-1/2)$.

2. Set the Window viewing variables in order to view the graph.

   Press [WINDOW], row 1, column 2. Set Xmin to -3, being sure to use the grey negation key. Set Xmax to 3; Xscl to 1; Ymin to -0.2, again being sure to use the grey negation key. Set Ymax to 0.5; Yscl to 1; Xres to 1. These settings are illustrated in Figure 1.10

3. View the graph.

   Press [GRAPH], row 1, column 5. The graph of the normal curve is shown in Figure 1.11.

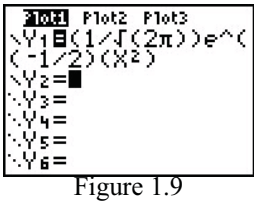


Figure 1.9

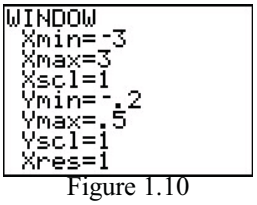


Figure 1.10

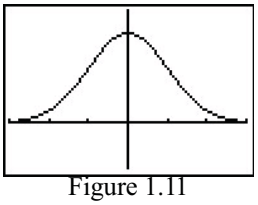


Figure 1.11

4. Set the graph window to standard viewing and clear the function .

   Press [ZOOM] and select 6: ZStandard to restore the default graph window settings. Press [Y=] and press [CLEAR] to remove the function.

## 1.4 Statistics

**Keystrokes Introduced**

1. [STAT] displays the Stat menu.
2. [STAT]>CALC displays the [STAT]>CALC menu.
3. [2nd] [DISTR] displays the distributions menu.
4. [VARS] displays the VARS menu.
5. [DISTR]>DISTR displays menu options for calculating values of common probability distributions.

6. DISTR >DRAW displays menu options for shading areas under a probability distribution function.
7. 2nd STAT PLOT displays statistical plot options.
8. STO→ stores values to a list or a single value to a variable.
9. ZOOM >ZoomStat redefines the viewing window so that all statistical data points are displayed.
10. TRACE may be used to trace a plot of statistical data.
11. ClrList clears from memory the elements of one or more listnames.
12. 2nd A-LOCK sets alpha lock on; ALPHA turns alpha lock off when alpha lock is on.
13. STAT >SetUpEditor clears the list editor and restores the built in list L!-l6.

The TI-83 Plus SE and TI-84 Plus SE have several functions for analyzing data. Many of these functions are contained in the STAT >CALC and STAT >TESTS menu options. The STAT key is located on the 3rd row, 3rd column. These menus are shown in Figure 1.12 and Figure 1.13. These functions provide summary statistics, regression lines, confidence intervals, hypothesis tests, and analysis of variance.

Other statistical functions are contained in the 2nd DISTR menu, located on the 4th row, 4th column above VARS. . DISTR >DISTR provide menu options for calculating values of common probability distribution functions, and is shown in Figure 1.14; DISTR >DRAW provide menu options for shading areas under a probability distribution function, and is shown in Figure 1.15.

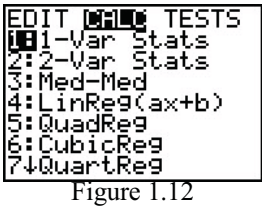


Figure 1.12

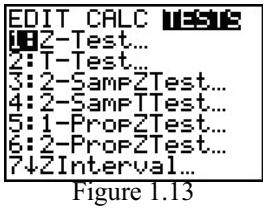


Figure 1.13

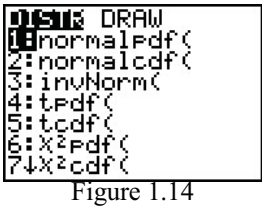


Figure 1.14

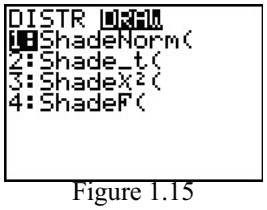


Figure 1.15

### Plotting Statistical Data

You can plot statistical data by selecting 2nd STAT PLOT, located directly over Y=.The 2nd STAT PLOT menu options provides access to statistical plot op-

tions and the capability of turning on/off all statistical plots, as shown in Figure 1.16. One, two, or all three statistical plots may be displayed on the screen simultaneously. The TI-83 Plus SE and TI-84 Plus SE can display a scatter plot, xyLine, histogram, modified box plot, regular box plot, and normal probability plot.

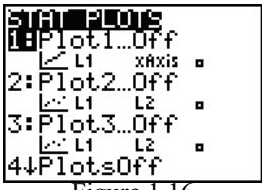


Figure 1.16

## Lists

Lists represent a set of observations. A list may contain up to 999 numerical values and is the principal way to store data for analysis. Many of the built-in statistical functions and programs operate on data sorted in a list or lists.The TI-83 Plus SE and TI-84 Plus SE have six list names in memory: L1, L2, L3, L4, L5, and L6. The list names L1 through L6 are on the keyboard above the numeric keys 1 through 6. To paste one of these names to a valid screen, press 2nd, and then press the appropriate key. L1 through L6 are stored in stat list editor columns 1 through 6 when you reset memory. Lists may also be created with a descriptive name. The name must be a string of up to 5 characters. The first letter must be a letter which may be folowed by letters, numbers, or $\theta$. The number of lists is limited by available memory. Lists may be created on the home screen, or in the STAT list editor.

**Example 1.3** Here are the weights (in pounds) of 18 men who were on the crew teams at *Oxford* and *Cambridge* universities (*The Independent*, March 31, 1992), also Hand, D. J. et al., 1994, p337.): (Utts/Heckard, Statistical Ideas and Methods, p27)

| | | | | | |
|---|---|---|---|---|---|
| *Cambridge* | 188.5 | 183.0 | 194.5 | 185.0 | 214.0 |
| | 203.5 | 186.0 | 178.5 | 109.0 | |
| *Oxford* | 186.0 | 184.5 | 204.0 | 184.5 | 195.5 |
| | 202.5 | 174.0 | 183.0 | 109.5 | |

Follow these steps to create two lists.

1. Create list L1 on the home screen.

   On the home screen, curly braces ({}) are used to enclose lists. Numbers are seperated by commas. Enter the weights for Cambridge within curly braces seperated by commas, as shown in Figure 1.17. Store the list by using the keystrokes STO→; 2nd L1; ENTER, storing the data in list L1. After pressing ENTER, the contents of the list are displayed on the home screen. Note that spaces rather than commas seperate values in a displayed list. You

may use the left and right arrow keys, ◄ and ►,to scroll through the list.

2. Create list L2 using the STAT list editor.

   Press STAT ENTER to select the STAT list editor. Note that the weights for Cambridge are displayed in list L1. Place the cursor on list L2 row 1 to make L2(1) the active list row, as shown in Figure 1.18. Enter the weights for Oxford pressing ENTER after each entry. The list is partially entered in Figure 1.19. Press 2nd QUIT to quit the STAT editor.

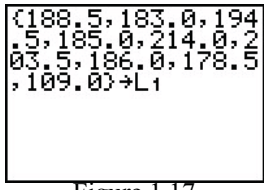


Figure 1.17

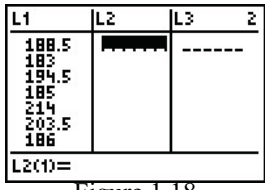


Figure 1.18

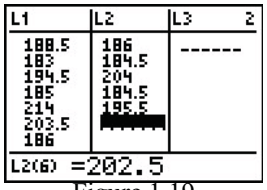


Figure 1.19

3. Plot the statistical data by creating modified box plots for the weights of the crew teams at *Oxford* and *Cambridge* universities.

   Press 2nd STAT PLOT accessing the StatPlot menu, as shown in Figure 1.20.

   Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified box plot. Press ENTER. Use the down arrow key to select L1 as the list, 2nd L1, as shown in Figure 1.21.

   Use the up arrow key to place the cursor on Plot2. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified box plot. Press ENTER. Use the down arrow key to select L2 as the list, 2nd L2, as shown in Figure 1.22.

   Press ZOOM, ZoomStat to view the graph, as shown in Figure 1.23.

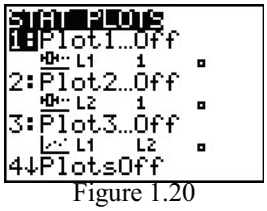


Figure 1.20

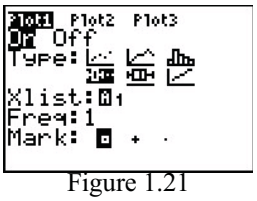


Figure 1.21

Figure 1.22

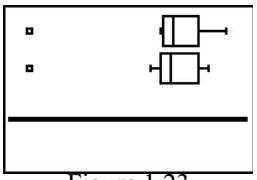


Figure 1.23

4. Identify outliers.

   Press the TRACE key and the left arrow key to identify an outlier (109.0) in list L1. Use the down arrow key and the left arrow key to identify an outlier (109.5) in list L2.

5. Turn off all plots and return the graph window to standard viewing.

   Press 2nd STAT PLOT, selecting PlotsOff and press ENTER. Press ZOOM and select 6: ZStandard to restore the default graph window settings.

6. Save list L1 as CAMBR and list L2 as OXFRD.

   Press 2nd L1 STO→ 2nd A-LOCK and type CAMBR; press ENTER.
   Press 2nd L2 STO→ 2nd A-LOCK and type OXFRD; press ENTER.

These outliers indicate that the last weight given in each list is very different from the others. In fact, those two men were the coxswains for their teams, while the other men were the rowers.

### Clearing Lists

To clear all of the entries in a list, or lists, press STAT, selecting 4: ClrList. Press the appropriate key. L1 through L6, as shown in Figure 1.24. To clear all lists, press STAT, selecting 4: ClrList. Press the appropriate keys, seperating each list name by a comma, as shown in Figure 1.25.

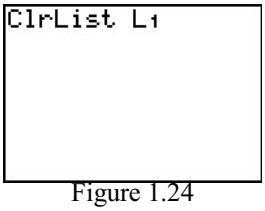


Figure 1.24

Figure 1.25

### Dispaying Lists

The menu option of STAT, SetUpEditor used without any arguments clears the list editor and restores the built in lists L1-L6. SetUpEditor followed by a sequence of up to 20 lists replaces the stat list editor with the new sequence of lists.

# Chapter 2
# Turning Data Into Information

## 2.1 Introduction

In this chapter, you will learn how to create simple summaries and pictures from various kinds of raw data.

After reading this chapter you should be able to:

1. Change frequencies to a percentage falling into each category.
2. Create a bar chart for a single categorical variable.
3. Create a bar chart displaying two categorical variables.
4. Obtain the five-number summary for quantitative data.
5. Plot statistical data by creating a histogram for a quantitative variable.
6. Create comparative boxplots for quantitative variables.
7. Draw a histogram with s superimposed normal curve.
8. Calculate the variance and standard deviation for a small data set.

## 2.2 Raw Data

Raw data is a term used for numbers and category labels that have been collected but have not yet been processed in any way. For example, here is a list of questions asked in a large statistics class and the "raw data" given by one of the students:

| | Question | Raw Data |
|---|---|---|
| 1. | What is your sex (m = male,f = female)? | m |
| 2. | How many hours did you sleep last night? | 5 hours |
| 3. | Randomly pick a letter-S or Q. | S |
| 4. | What is your height in inches? | 67 inches |
| 5. | Randomly pick a number between 1and 10. | 3 |
| 6. | What's the fastest you've ever driven a car (mph)? | 110 mph |
| 7. | What is your right handspan In centimeters? | 21.5 cm |
| 8. | What is your left handspan in centimeters? | 21.5 cm |

## 2.3 Types of Variables

Different types of summaries are appropriate for different types of variables. It makes sense, for example, to calculate the average number of hours of sleep last night for the members of a group, but it doesn't make sense to calculate the average sex (male, female) for the group. For gender data, it makes more sense to determine

the proportion of the group that's male and the proportion that's female.
We learned in a previous section that a variable is a characteristic that differs from one individual to the next. A variable may be a categorical characteristic, like a person's sex, or a numerical characteristic, like hours of sleep last night.

| | Example |
|---|---|
| Raw data from categorical variables consist of group or category names that don't necessarily have a logical ordering. | eye color |
| Categorical variables for which the categories have a logic ordering are called ordinal variables. | highest degree earned |
| Raw data from quantitative variables consist of numerical values taken on each individual | height in inches |

TI calculators allow only for numerical values to be used in a statistical analysis. For example, the text "Male" or "Female" can not be used for the "Sex" variable in the *PennState1* worksheet. Neither can we use the letters "M" or "F" since these letters are replaced by the value stored in memory for the "M" and "F" variables in the calculator.
The solution to the problem is to assign a unique numerical code for each value of the variable. In this case, you might code "Male = 0" and "Female = 1" on the TI calculator.

Values of the other categorical categorical variables ("SQpick" and "Form") in the *PennState1* worksheet could also be coded. For example, you might code "S = 0" and "Q = 1". Other numerical values could also be used.

The quantitative variables in the *PennState1* worksheet: (Hours of sleep the previous night, Choice of either S or Q, Reported height, inches , "Random" pick of a number between 1 and 10, Fastest speed ever driven, mph, Measured stretched right handspan, cm, Measured stretched left handspan, cm ) can be handled by the TI calculator without coding.

## 2.4 Summarizing One or Two Categorical Variables

### Numerical Summaries

To summarize a categorical variable, the first step is to count how many individuals fall into each possible category. Percents usually are more informative than counts so the second step is to calculate the percent in each category. These two easy steps can also be used to summarize a combination of two categorical variables.

**Keystrokes Introduced**

1. [2nd] [LIST] >MATH>sum( returns the sum of the elements within a list.
2. [ZOOM] >ZStat redefines the viewing window so that all statistical data points

are displayed.

3. [2nd] [STAT PLOT] accesses the StatPlot menu.
4. [2nd] [DRAW] > Text( draws text on a graph screen.
5. [STAT] >CALC> 1: 1-VarStats analyzes data for one quantitative variable.
6. [2nd] [LIST] [►], OPS. Select 1: SortA( sorts elements of a list in ascending order.

**Example 2.1 Seatbelt Use by 12thGraders**

How often do you wear a seatbelt when driving a car? This is one of many questions asked in a biennial nationwide survey of American high school students. The survey, conducted as part of a federal program called the Youth Risk Behavior Surveillance System (YRBSS), is sponsored and organized by the U.S. Centers for Disease Control (CDC). Survey questions concern potentially risky behaviors such as cigarette smoking, alcohol use, and so on. For the question about seatbelt use when driving, possible answers were Always, Most times, Sometimes, Rarely, and Never. An additional choice allowed respondents to say they don't drive, which often was the case because many survey participants were under the minimum legal driving age. Table 2.1 summarizes responses in the 2003 survey given by 12thgrade students who said they drive.

| Response | Count |
|---|---|
| Always | 1686 |
| Most times | 578 |
| Sometimes | 414 |
| Rarely | 249 |
| Never | 115 |

Table 2.1

**Follow these steps** to determine the percentage of students falling into each category.

1. Clear any data from lists L1 and L2.

   Press [STAT] [ENTER] to select the [STAT] list editor. Place the cursor at the top of list L1. Press [CLEAR] followed by the down arrow key, ▼ to clear any data from list L1. Place the cursor at the top of list L2. Press [CLEAR] followed by the down arrow key, ▼ to clear any data from list L2.

2. Enter the data.

   Place the cursor on list L1 row 1 to make L1(1) the active list row, as shown in Figure 2.1. Enter the counts for the responses pressing [ENTER] after each entry. The list is entered in Figure 2.2.

3. Enter an expression to determine the percentage of students falling into each

category.

Move the cursor to the top of list L2. With the cursor at the top of list L2 type [2nd] L1 ÷ [2nd] [LIST] [◄], selecting sum( , and press [ENTER]. Type [2nd] L1 and a ). These steps are reflected in Figures 2.3 and 2.4. Press [ENTER] to evaluate the expression.

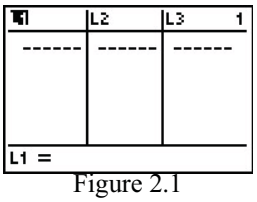


Figure 2.1

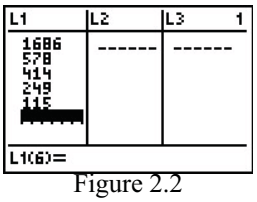


Figure 2.2

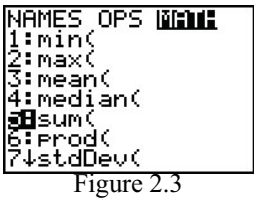


Figure 2.3

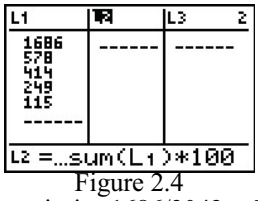


Figure 2.4

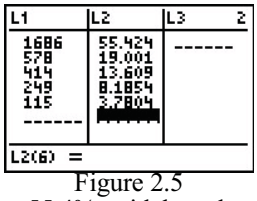


Figure 2.5

Notice that a majority, 1686/3042= .554 or 55.4%, said they always wear a seatbelt when driving, while just 115/3042= .032 or 3.2% said they never wear a seatbelt. Because 55.4% said they always wear a seatbelt, we can calculate the percent who don't always wear a seatbelt as 100%-55.4% =44.6% . Alternatively, the percent saying they don't always wear a seatbelt could be detennined as 19.0% + 13.6% + 8.2% + 3.8%, the sum of the percents for all categories other than Always.

## Frequency and Relative Frequency

Frequency is a synonym for the count of how many observations fall into a category. The proportion or percent in a category is a type of a relative frequency, the count in a category relative to the total count over all categories. A frequency distribution for a categorical variable is a listing of all categories along with their frequencies (counts). A relative frequency distribution is a listing of all categories along with their relative frequencies (given as proportions or percents, for example). It is commonplace to give the frequency and relative frequency distributions together, as was done in Table 2.1

## Visual Summaries for Categorical Variables

There are two simple visual summaries used for categorical data:

a. Pie charts are useful for summarizing a single categorical variable if there are not too many categories. Unfortunately, pie charts are not built-in to the TI-83 Plus SE nor the TI-84 Plus SE.

b. Bar graphs are useful for summarizing one or two categorical variables and are particularly useful for making comparisons when there are two categorical variables.

Both of these simple graphical displays are easy to construct and interpret, as the examples in the text demonstrate.

**Example 2.3 Random Numbers** Question 5 in the class survey described in Section 2.1 asked students to "Randomly pick a number between 1and 10." The pie chart shown in Figure 2.1 of the text illustrates that the results are not even close to being evenly distributed across the numbers. Notice that almost 30%of the students chose 7 while only just over 1% chose the number l. The data is displayed as an ungrouped frequency distribution in Table 2.2.

| Random Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Percent | 1 | 4.7 | 11.6 | 11.0 | 9.5 | 12.1 | 29.5 | 10 | 7.4 | 3.2 |
| Frequency | 2 | 9 | 22 | 21 | 18 | 23 | 56 | 19 | 14 | 6 |

Table 2.2

**Follow these steps** to create a bar chart for the categorical variable "random number".

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6, as shown in Figure 2.6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the data for the categorical variable "random number" in list L1.

   Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the data: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 pressing [ENTER] after each entry. Place the cursor on list L12 row 1 to make L2(1) the active list row. Enter the frequencies: 2, 9, 22, 21, 18, 23, 56, 19, 14, 6 in L2 pressing [ENTER]

after each entry, as shown in Figure 2.7.

Figure 2.6

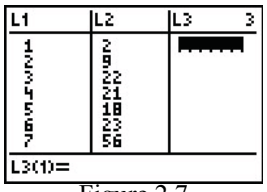


Figure 2.7

3. Plot the statistical data by creating a bar chart for the categorical variables "random number".

   Press 2nd STAT PLOT accessing the StatPlot menu.

   Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram (bar chart). Press ENTER. Use the down arrow key to select list L1 as the list, 2nd L1. Use the down arrow key to enter list L2 as the Freq: 2nd L2. The settings for Plot 1 are shown in Figure 2.8.

4. View the graph.

   Press ZOOM, 9: ZoomStat to view the graph, as shown in Figure 2.9.

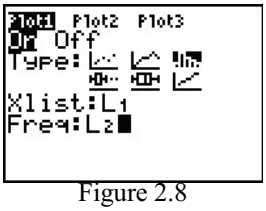


Figure 2.8

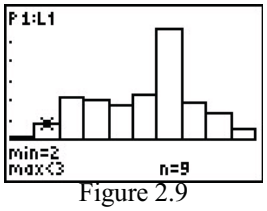


Figure 2.9

**Example 2.4 Myopia** A survey of 479 children found that those who had slept with a nightlight or in a fully lit room before the age of 2 had a higher incidence of nearsightness (myopia) later in childhood (*Sacramento Bee*, May 13, 1999, pp. A1, A18). The raw data for each child consisted of two categorical variables, each with three categories. Table 2.2 gives the categories and the number of children falling into each combination of them.

The patten in Table 2.2 is striking. As the amount of sleeptime light inceases, the incidence of myopia also increases. However this study does not prove that sleeping with light actually caused myopia in children. There are other possible explanations. For example, myopia has a genetic component, so those children whose parents have myopia are more likely to suffer from it themselves. Maybe nearsighted parents are more likely to proviode light while their children are

sleeping.

| Slept with: | No Myopia | Myopia | High Myopia | Total |
|---|---|---|---|---|
| Darkness | 155 (90%) | 15 (9%) | 2 (1%) | 172 |
| Nightlight | 153 (66%) | 72 (31%) | 7 (3%) | 232 |
| Full Light | 34 (45%) | 36 (48%) | 5 (7%) | 75 |
| Total | 342 (71%) | 123 (26%) | 14 (3%) | 479 |

Table 2.2

**Follow these steps** to create a bar chart for the categorical variables. You will create a clustered bar chart displayed in percentages of the row totals for each of the categorical variables.

5. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6, as shown in Figure 2.10. Press [ENTER] to execute the command.

6. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter codes for categorical variable "Slept with" in odd.numbered lists.

      Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter 1, 2, 3 pressing [ENTER] after each entry.
      Place the cursor on list L3 row 1 to make L3(1) the active list row. Enter 5, 6, 7 pressing [ENTER] after each entry.
      Place the cursor on list L5 row 1 to make L5(1) the active list row. Enter 9, 10, 11 pressing [ENTER] after each entry.
   b. Enter the percentages, as whole numbers, for the categorical variable "Myopia" in even numbered lists.

      Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the percentages: 90, 9, 1 pressing [ENTER] after each entry.
      Place the cursor on list L4 row 1 to make L4(1) the active list row. Enter the percentages: 66, 31, 3 pressing [ENTER] after each entry.
      Place the cursor on list L6 row 1 to make L6(1) the active list row. Enter

the percentages: 45, 48, 7 pressing [ENTER] after each entry.
The results of the data entry process are shown in Figures 2.11 and 2.12.

Figure 2.10

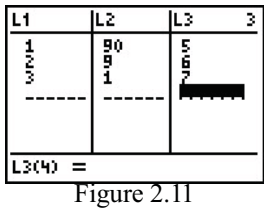


Figure 2.11

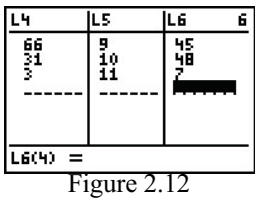


Figure 2.12

7. Plot the statistical data by creating a clusted bar chart for the categorical variables "Slept with" and "Myopia".

Press [2nd] [STAT PLOT] accessing the StatPlot menu.

(i) Press [ENTER], selecting Plot 1. Place the cursor on ON and press [ENTER]. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram. Press [ENTER]. Use the down arrow key to select L1 as the Xlist, [2nd] [L1]. Use the down arrow key to select L2 as the Freq:, [2nd] [L2]. The settings for Plot 1 are shown in Figure 2.13.

(ii) Use the up arrow key to place the cursor on Plot2. Place the cursor on ON and press [ENTER]. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram. Press [ENTER]. Use the down arrow key to select L3 as the Xlist, [2nd] [L3]. Use the down arrow key to select L4 as the Freq:, [2nd] [L4]. The settings for Plot 2 are shown in Figure 2.14.

(iii) Use the up arrow key to place the cursor on Plot3. Place the cursor on ON and press [ENTER]. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram. Press [ENTER]. Use the down arrow key to select L5 as the Xlist, [2nd] [L5]. Use the down arrow key to select L6 as the Freq:, [2nd] [L6]. The settings for Plot 3 are shown in Figure 2.15.

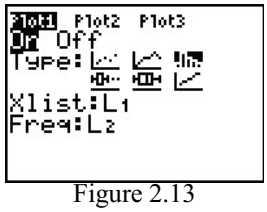


Figure 2.13

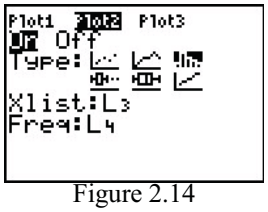


Figure 2.14

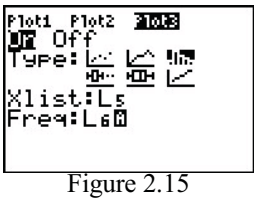


Figure 2.15

8. Set the Window viewing variables in order to view the graph.

Press [WINDOW], row 1, column 2. Set Xmin to 1. Set Xmax to 12; Xscl to 1; Ymin to -10, being sure to use the grey negation key. Set Ymax to 105; Yscl

to 10; Xres to 1. These settings are illustrated in Figure 2.16

9. View the graph.

   Press GRAPH to view the graph, as shown in Figure 2.16.

Figure 2.16

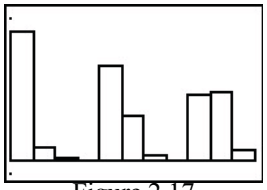


Figure 2.17

10. Optional: Add text to the histogram (bar chart).

    Press 2nd DRAW, selecting 0: Text from the DRAW menu, as shown in Figure 2.18. Use the arrow keys to position the cursor. Press 2nd A-LOCK to type the labels. You may have to select 2nd DRAW, selecting 1: ClrDraw from the DRAW menu and GRAPH to attempt once again to position the labels to your satisfaction. The finished graph is displayed in Figure 2.19.

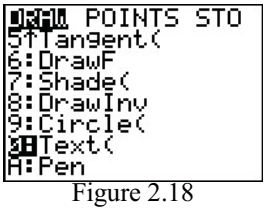


Figure 2.18

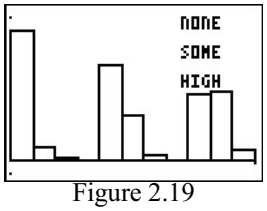


Figure 2.19

The first cluster on the left of the clustered bar chart displays the category "Darkness" of the "Slept with" variable. The heights of the bars indicate relative frequencies of 90%, 9%, and 1%. The middle cluster of the clustered bar chart displays the category "Nightlight" of the "Slept with" variable. The heights of the bars indicate relative frequencies of 66%, 31%, and 3%. The third cluster from the left of the clustered bar chart displays the category "Full light" of the "Slept with" variable. The heights of the bars indicate relative frequencies of 45%, 48%, and 7%.

11. Turn off all plots and return the graph window to standard viewing.

    Press 2nd STAT PLOT, selecting PlotsOff and press ENTER. Press ZOOM and select 6: ZStandard to restore the default graph window settings.

These outliers indicate that the last weight given in each list is very different from the others. In fact, those two men were the coxswains for their teams, while the other men were the rowers.

## 2.4 - (continued) Interesting Features of Quantitative Data

Looking at a long, disorganized list of data values is about the same as looking at a scrambled set of letters. To begin finding the information in quantitative data, we have to organize it using visual displays and numerical summaries. In this section we focus on interpreting the main features of quantitative variables. More specific details will be given in the following sections.

**Example 2.5 Right Handspans.** Table 2.3 displays the raw data for the right handspan measurements (in centimeters) made in the student survey described in Section 2.1 of the text. The measurements are listed separately for males and females, but are not organized in any other way. Imagine that you know a female whose stretched right handspan is 20.5 em. Can you see how she compares to the other females in Table 2.3? That probably will be hard because the list of data values is disorganized.
We will organize the handspan data in Table 2.3 using a five-number summary, which consists of the median, the quartiles (roughly, the medians of the lower and upper halves of the data), and the extremes (high, low).

Males (87 students)
21.5, 22.5, 23.5, 23.0, 24.5, 23.0, 26.0, 23.0, 21.5, 21.5,
24.5, 23.5, 22.0, 23.5, 22.0, 22.0, 24.5, 23.0, 22.5, 19.5,
22.5, 22.0, 23.0, 22.5, 20.5, 21.5, 23.0, 22.5, 21.5, 25.0,
24.0, 21.5, 21.5, 18.0, 20.0, 22.0, 24.0, 22.0, 23.0, 22.0,
22.0, 23.0, 22.5, 25.5, 24.0, 23.5, 21.0, 25.5, 23.0, 22.5,
24.0, 21.5, 22.0, 22.5, 23.0, 18.5, 21.0, 24.0, 23.5, 24.5,
23.0, 22.0, 23.0, 23.0, 24.0, 24.5, 20.5, 24.0, 22.0, 23.0,
21.0, 22.5, 21.5, 24.5, 22.0, 22.0, 21.0, 23.0, 22.5, 24.0,
22.5, 23.0, 23.0, 23.0, 21.5, 19.0, 21.5

Females (103 students)
20.00, 19.00, 20.50, 20.50, 20.25, 20.00, 18.00, 20.50, 22.00,
20.00, 21.50, 17.00, 16.00, 22.00, 22.00, 20.00, 20.00, 20.00,
20.00, 21.70, 22.00, 20.00, 21.00, 21.00, 19.00, 21.00, 20.25,
21.00, 22.00, 18.00, 20.00, 21.00, 19.00, 22.50, 21.00, 20.00,
19.00, 21.00, 20.50, 21.00, 22.00, 20.00, 20.00, 18.00, 21.00,
22.50, 22.50, 19.00, 19.00, 19.00, 22.50, 20.00, 13.00, 20.00,
22.50, 19.50, 18.50, 19.00, 17.50, 18.00, 21.00, 19.50, 20.00,
19.00, 21.50, 18.00, 19.00, 19.50, 20.00, 22.50, 21.00, 18.00,
22.00, 18.50, 19.00, 22.00, 12.50, 18.00, 20.50, 19.00, 20.00,
21.00, 19.00, 19.00, 21.00, 18.50, 19.00, 21.50, 21.50, 23.00,
23.25, 20.00, 18.80, 21.00, 17.00, 21.00, 20.00, 20.50, 20.00,
19.50, 21.00, 21.00, 20.00
Table 2.3

**Follow these steps** to obtain the five-number summaries for females and males.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the "Stretched Right Handspans (cm) " of the 190 College students in lists L1 and L2.

      Enter the "Stretched Right Handspans (cm) " for the Males (87 students) in list L1. Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter 21.5, 22.5, 23.5, etc. pressing [ENTER] after each entry.
      Enter the "Stretched Right Handspans (cm) " for the Females (103 students) in list L2. Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter 20, 19, 20.5, etc. pressing [ENTER] after each entry.

      The data in lists L1 and L2 are displayed in Figure 2.20

3. Obtain the five-number summaries for females and males.

   Press [STAT] [▶] to obtain the [STAT] CALC menu, as shown in Figure 2.21.

   a. Select 1: 1-Var Stats and press [ENTER]. Press [2nd] [L1] to select the "Stretched Right Handspans (cm) " of the males. Press [ENTER]. Use the down arrow key, [▼], five times. The output from the TI calculator is displayed in Figure 2.22.
   b. Select 1: 1-Var Stats and press [ENTER]. Press [2nd] [L1] to select the "Stretched Right Handspans (cm) " of the males. Press [ENTER]. Use the down arrow key, [▼], five times. The output from the TI calculator is displayed in Figure 2.23.

4. Save list L1 as MRSN1 and list L2 as FRSN1.

   Press [2nd] [L1] [STO→] [2nd] [A-LOCK] and type MRSN [ALPHA] 1; press

ENTER. Press 2nd L2 STO→ 2nd A-LOCK and type FRSN ALPHA 1; press ENTER.

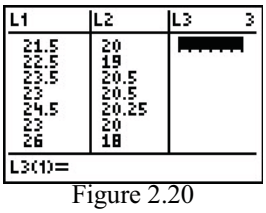


Figure 2.20

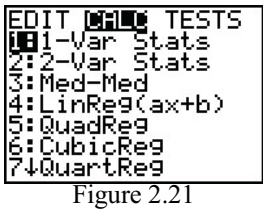


Figure 2.21

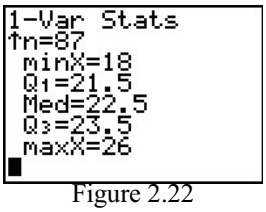


Figure 2.22

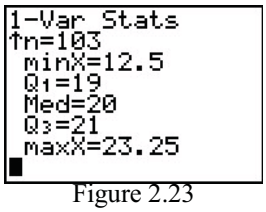


Figure 2.23

Remember that the five-number summary approximately divides the dataset into quarters. For example, about 25% of the female handspan measurements are between 12.5 and 19.0 centimeters, about 25% are between 19 and 20 em, about 25% are between 20 and 21 em, and about 25 % are between 21 and 23.25 em. The five-number summary gives us a good idea of where our imagined female with the 20.5 centimeter handspan fits into the distribution of handspans for females. She's in the third quarter of the data, slightly above the median (the middle value).

## 2.6 Pictures for Quantitative Data

There are three similar types of pictures that are used to represent quantitative variables, all of which are valuable for assessing center, spread, shape, and outliers. Histograms are similar to bar graphs and can be used for any number of data values, although they are not particularly informative when the sample size is small. Stem-and-Leaf plots and dotplots present all individual values, so for very large datasets they are more cumbersome than histograms. A fourth kind of picture, called a boxplot or box-andwhisker plot, displays the information given in a five-number summary. It is especially useful for comparing two or more groups and for identifying outliers. The TI-83 Plus SE and the TI-84 Plus SE are well suited for displaying histograms, Stem-and-Leaf plots and boxplots. The TI-83 Plus SE and the TI-84 Plus SE do not have build in features for creating dotplots. We will begin by creating a histogram of women's right handspans.

**Example 2.5 Right Handspans.** Table 2.3 displays the right handspan measurements (in centimeters) made in the student survey described in Section 2.1 of the text. The measurements are listed separately for males and females. Recall that the right handspan measurements for the females are stored in the list FRSN1.

**Follow these steps** to obtain the histogram of right handspans for females.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the "Stretched Right Handspans (cm) " of the females in lists L1.

      Place the cursor at the top of list L1. Press [2nd] [LIST], selecting the list FRSN1, as shown in Figure 2.24. Press [ENTER] to drive the data into the working list L1. The data from the list FRSN1 is displayed in list L1, as shown in Figure 2.25.

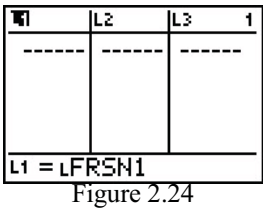


Figure 2.24

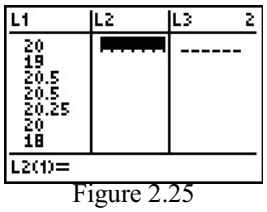


Figure 2.25

3. Plot the statistical data by creating a histogram of the right handspan measurements for the females.

   Press [2nd] [STAT PLOT] accessing the StatPlot menu.

   (i) Press [ENTER], selecting Plot 1. Place the cursor on ON and press [ENTER]. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram. Press [ENTER]. Use the down arrow key to select L1 as the list, [2nd] [L1]. Use the down arrow key to enter 1 as the Freq:. The settings for Plot 1 are shown in Figure 2.26.

4. View the graph.

   Press [ZOOM], 9: ZoomStat to view the graph, as shown in Figure 2.27.

Figure 2.26

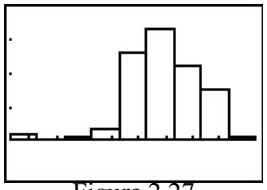


Figure 2.27

5. Turn off all plots and return the graph window to standard viewing.

   Press 2nd STAT PLOT, selecting PlotsOff and press ENTER. Press ZOOM and select 6: ZStandard to restore the default graph window settings.

The histogram shows the distribution of the data, the pattern of how often the various measurements occurred. The histogram is useful for assessing the location, spread, and shape of a distribution and may be useful for detecting outliers. Notice that the values are "centered" around 20 em, which is the median value. There are two possible outlier values that are low compared to the bulk of the data that are evident in the histogram. Except for those values, the handspans have a range of about 7 em, extending from about 16 to 23 em. They tend to be "clumped" around 20 and taper off toward 16 and 23.

**Example 2.5 Continued. Right Handspans.** Table 2.3 displays the right handspan measurements (in centimeters) made in the student survey described in Section 2.1 of the text. The measurements are listed separately for males and females. Recall that the right handspan measurements for the females are stored in the list FRSN1 and the right handspan measurements for the males are stored in the list MRSN1.

**Follow these steps** to obtain the boxplot of right handspans for females and males.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press Y= and press CLEAR to remove all functions For each line that is not blank, place the cursor on the function and press CLEAR Press 2nd QUIT.

   b. Clear all lists in the Stat editor.

      Press STAT, selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press ENTER to execute the command.

2. Enter data using the STAT list editor.

   Press STAT ENTER to select the STAT list editor.

   a. Enter the "Stretched Right Handspans (cm) " of the females in lists L1 and the "Stretched Right Handspans (cm) " of the males in lists L2.

   Place the cursor at the top of list L1. Press 2nd LIST, selecting the list FRSN1, as shown in Figure 2.28. Press ENTER to drive the data into the working list L1. Place the cursor at the top of list L2. Press 2nd LIST, selecting the list MRSN1, as shown in Figure 2.29. Press ENTER to drive the data into the working list L1. The data from the list FRSN1 and MRSN1 is displayed in list L1 and list L2, as shown in Figure 2.30.

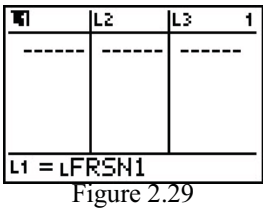


Figure 2.29

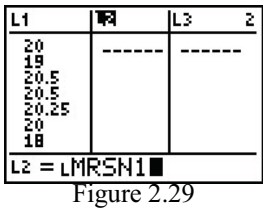


Figure 2.29

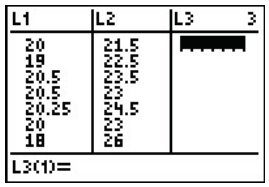


Figure 2.30

3. Plot the statistical data by creating comparative boxplots of the right handspan measurements for the females and males.

   Press 2nd STAT PLOT accessing the StatPlot menu.

   (i) Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select L1 as the list, 2nd L1. Use the down arrow key to enter 1 as the Freq:. The settings for Plot 1 are shown in Figure 2.31.

   (ii) Use the up arrow key to place the cursor on Plot2. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified box plot. Press ENTER. Use the down arrow key to select L2 as the

list, 2nd L2. Use the down arrow key to enter 1 as the Freq:. The settings for Plot 2 are shown in Figure 2.32.

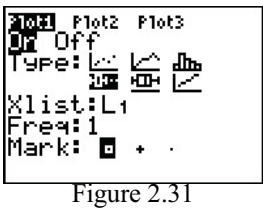


Figure 2.31

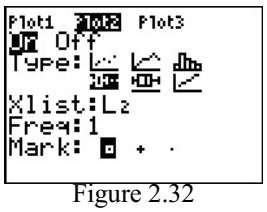


Figure 2.32

4. View the graph.

Press ZOOM, 9: ZoomStat to view the graph, as shown in Figure 2.33.

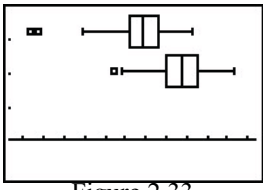


Figure 2.33

5. Turn off all plots and return the graph window to standard viewing.

Press 2nd STAT PLOT, selecting PlotsOff and press ENTER. Press ZOOM and select 6: ZStandard to restore the default graph window settings.

The comparative boxplots compare the spans of the right hands of males and females. For each group, the box covers the middle 50% of the data, and the line within a box marks the median value. With the exception of possible outliers, the lines extending from a box reach to the minimum and maximum data values. Possible outliers are marked with an square.

## 2.7 Numerical Summaries of Quantitative Variables

We discussed the interesting features of a quantitative dataset in Section 2.4 of the text, and in Section 2.5 of the text we learned how to look for them using use visual displays of the data. In this section we learn how to compute numerical summaries of these features for quantitative data.

### Quartiles and Five-Number Summaries

A simple way to find the quartiles is to split the ordered values into the half that is below the median and the half that is above the median. The lower quartile (Ql) is the median of the data values ,that are below the median. The upper quartile (Q3) is the median of the data values that are above the median. These values are

called quartiles because, along with the median and the extremes, they approximately divide the ordered data into quarters.We will begin by creating a histogram of women's right handspans.

**Example 2.13 Fastest Speeds.** In Case Study 1.1 we summarized responses to the question "What's the fastest you've ever driven a car?' Table 2.4 displays the response of the 87 males surveyed.

110,109,90,140,105,150,120,110,110,90,115,95,145,
140,110,105,85,95,100,115,124,95,100,125,140,85,
120,115,105,125,102,85,120,110,120,115,94,125,80,
85,140,120,92,130,125,110,90,110,110,95,95,110,
105,80,100,110,130,105,105,120,90,100,105,100,120,
100,100,80,100,120,105,60,125,120,100,115,95,110,
101,80,112,120,110,115,125,55,90

Table 2.4

**Follow these steps** to obtain the five-number summary for the 87 speeds.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the "Fastest Speeds" in lists L1.

      Enter the "Fastest Speeds" for the 87 students in list L1. Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter 110, 109, 90, etc. pressing [ENTER] after each entry. After entering all of the data, select

2nd QUIT.

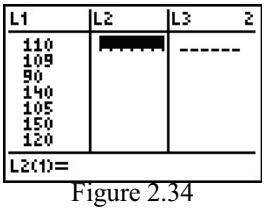


Figure 2.34

Figure 2.35

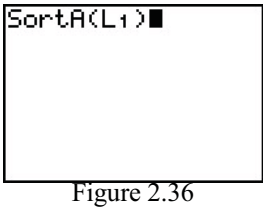


Figure 2.36

b. Save list L1 as MFST1.

Press 2nd L1 STO→ 2nd A-LOCK and type MFST ALPHA 1; press ENTER, as shown in Figure 2.35.

c. Sort the data in ascending order.

Press 2nd LIST ▶, OPS. Select 1: SortA(, pressing ENTER. Press 2nd L1 ) and press ENTER, as shown in Figure 2.36.

d. Examine the data set.

Press STAT ENTER to select the STAT list editor. The sorted data now appears in list 1, as shown in Figure 2.37.

Use the down arrow key, ▼, to locate the 44th value in the list, as shown in Figure 2.38.

The median is the middle value in an ordered list, so for 87 values, the median is the (87 + 1)/2 =88/2 =44th value in the list. The 44th value is 110, and this value is shown in bold in the data list, as shown in Figure 2.38.

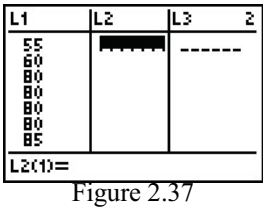


Figure 2.37

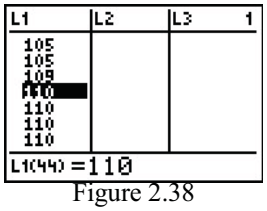


Figure 2.38

Aside from the middle value of 110, there were 43 values at or below 110, and another 43 values at or above 110. Notice that there are many responses of 110, which is why we are careful to say that 43 of the values are at or above the median.

There are 43 values on either side of the median. To find the quartiles, simply find the median of each of those sets of 43 values. The lower quartile is the (43 + 1)/2 = 22nd value from the bottom of the data.

Use the up arrow key, ▲, to locate the 22nd value in the list, as shown in Figure 2.39. The value of Q1 is 95.

Use the down arrow key, ▼, to locate the 22nd value from the top, as shown in Figure 2.40. The upper quartile is the 22nd value from the top;

the value of Q3 is 120.

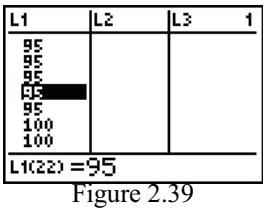


Figure 2.39

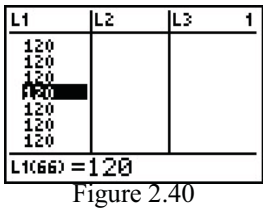


Figure 2.40

The median and quartiles divide the data into equal numbers of values but do not necessarily divide the data into equally wide intervals. For example, the lowest 1/4 of the males had responses ranging over the 40-mph interval from 55 mph to 95 mph, while the next 1/4 had responses ranging over only a l5-mph interval, from 95 to 110. Similarly, the third quarter had responses in only a 10-mph interval (110 to 120), while the top 1/4 had responses in a 30-mph interval (120 to 150). It is common to see the majority of values clumped in the middle and the remainder tapering off into a wider range.

e. Find the Summary Measures (mean, median, quartiles, low and high values, range and interquartile range).

Press [STAT] > CALC, selecting 1: 1Var Stats. Press [ENTER]. Press [2nd] [L1] and [ENTER], as shown in Figure 2.41. The results are shown in Figure 2.42. Use the down arrow key, [▼], five times to obtain the summary measure shown in Figure 2.43.

Figure 2.41

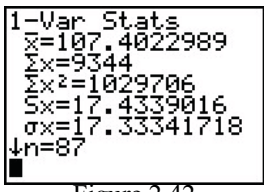


Figure 2.42

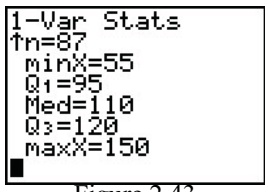


Figure 2.43

The calculator output indicates the mean = 107.4, miniumum =55, Q1 = 95, median = 110, Q3 = 120 and maximum = 150. The range is maximum - minimum= 150-55 = 95 and the interquartile range is Q3-Q1 = 120 - 95 = 25.

## 2.8 Features of Bell-Shaped Distributions

Nature seems to follow a predictable pattern for many kinds of measurements. Most individuals are clumped around the center, and the greater the distance a value is from the center, the fewer individuals have that value. Except for the two outliers at the lower end. that pattern is evident in the females' right handspan measurements, as shown in Example 2.5, Figure 2.27. If we were to draw a smooth curve onnecting the tops of the bars on a histogram with this shape, the smooth curve would resemble the shape of a bell.

Numerical variables that follow this pattern are said to follow a bell-shaped curve, or to be "bell-shaped." A special case of this distribution of measurements is so

common it is also called a normal disbibution or normal curve.

**Example 2.5 - Revisted - Women's Right Hand Spans.** Table 2.3 displays the raw data for the right handspan measurements (in centimeters) made in the student survey described in Section 2.1 of the text. The measurements are listed separately for males and females, but are not organized in any other way. In Example 2.5, you have saved the data for the males in list MRSN1 and the data for the females in list FRSN1.
We will draw a histogram of the women's right handspans, with a superimposed normal curve.

**Follow these steps** to draw the histogram of the women's right handspans, with a superimposed normal curve.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the "Stretched Right Handspans (cm) " of the females in list L1.

      Place the cursor at the top of list L1. Press [2nd] [LIST], selecting the list FRSN1, as shown in Figure 2.44. Press [ENTER] to drive the data into the working list L1. The data from the list FRSN1 is displayed in list L1, as shown in Figure 2.45.

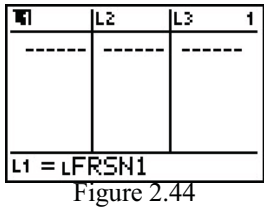


Figure 2.44

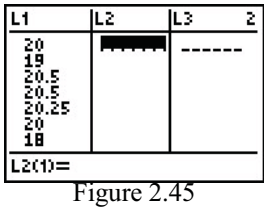


Figure 2.45

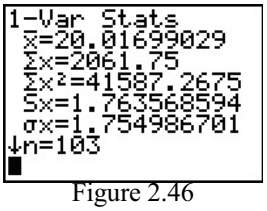


Figure 2.46

   b. Obtain the numerical summaries of the women's right handspans.

Press STAT ► to obtain the STAT CALC menu.

c. Select 1: 1-Var Stats and press ENTER. Press 2nd L1 to select the "Stretched Right Handspans (cm) " of the females. The output from the TI calculator is displayed in Figure 2.46.
Observe that the mean of the women's right handspans is 20.017 and the standard deviation is 1.764.

3. Set up the plot for the histogram of the right handspan measurements for the females.

Press 2nd STAT PLOT accessing the StatPlot menu.

(i) Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram. Press ENTER. Use the down arrow key to select L1 as the list, 2nd L1. Use the down arrow key to enter 1 as the Freq:. The settings for Plot 1 are shown in Figure 2.47.

4. Enter the function to superimpose the normal curve on the histogram.

Press Y=, row 1, column 1, to enter the function, as shown in Figure 1.9. Press $(18/1.764\sqrt{(2\pi)})e\hat{}((-1/2)(x-20.017)^2/1.764^2)$. Observe that the mean of the women's right handspans, 20.017 and the standard deviation, 1.764, are entered into the function to determine the y-values of the graph. The 18 is a scaling factor designed to make the plot of the histogram and the normal curve coincide. Other scaling factors can be explored. The left and right parentheses are located on row 6. Press 2nd $\pi$, $\pi$ is located on the 5th row, right column above the ^ key. Press 2nd $e$, $e$ is located on the 8th row, left column above the LN key. Be sure to use the grey negation key when you enter $(-1/2)$. The function is shown in Figure 2.48.

5. Set the Window viewing variables in order to view the graph.

Press WINDOW, row 1, column 2. Set Xmin to 11, Xmax to 27; Xscl to 1; Ymin to -5, being sure to use the grey negation key. Set Ymax to 31; Yscl to 1; Xres to 1. These settings are illustrated in Figure 2.49

6. View the graph.

Press GRAPH, to view the graph, as shown in Figure 2.50.

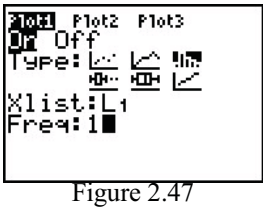


Figure 2.47

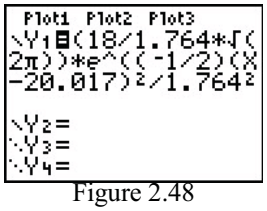


Figure 2.48

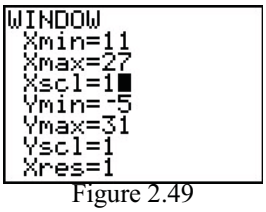


Figure 2.49

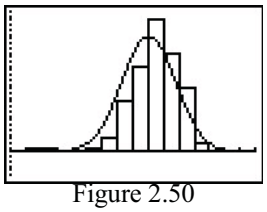


Figure 2.50

7. Turn off all plots and return the graph window to standard viewing.

   Press 2nd STAT PLOT, selecting PlotsOff and press ENTER. Press ZOOM and select 6: ZStandard to restore the default graph window settings.

8. Clear the function.

   Press Y=, and Press Y= and press CLEAR to remove all functions For each line that is not blank, place the cursor on the function and press CLEAR Press 2nd QUIT.

## The Concept of Standard Deviation

Because normal curves are so common in nature, a whole set of descriptive features has been developed that apply mostly to variables with that shape. In fact, two summary features uniquely determine a normal curve, so that if you know those two summary numbers, you can draw the curve precisely. The first summary number is the mean, and the bell shape is centered on that number. The second summary number is called the standard deviation, and it is a measure of the spread of the values.

You can think of the standard deviation as roughly the average distance values fall from the mean. Put another way, it measures variability by summarizing how far individual data values are from the mean.

The formula for calculating the standard deviation is a bit more involved than the conceptual interpretation just discussed. This is the first instance of a summary measure that differs based on whether the data represent a sample or an entire population. The version given here is appropriate when the dataset is considered to represent a sample from a larger population. The value of $s^2$, the squared standard deviation is called the (sample) **variance**. The formula for the (sample) variance is

$$s^2 = \frac{\sum(x - \overline{x})^2}{n - 1}$$

In practice, statistical software like Minitab, a spreadsheet program like Excel, or

a TI calculator typically is used to find the standard deviation lor a dataset. For situations where you have to calculate the standard deviation by hand, here is a step-bystep guide to the steps involved:

**Step 1:** Calculate $\overline{x}$, the sample mean.
**Step 2:** For each observation, calculate the difference between the data value and the mean.
**Step 3:** Square each difference calculated In step 2.
**Step 4:** Sum the squared differences calculated in step 3, and then divide this sum by $n - 1$.The answer for this step Is called the varIance.
**Step 5:** Take the square root of the variance calculated in step 4. The answer for this step is called the standard deviation.

**Example 2.18 Calculating a Standard Deviation.** You will calculate the standard deviation of the four pulse rates 62, 68, 74, 76.

**Follow these steps** to calculate the standard deviation.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the four pulse rates 62, 68, 74, 76 in list L1.

   Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter 62, 68, 74, 76 pressing [ENTER] after each entry. The data is displayed in

list L1, as shown in Figure 2.45. Press 2nd QUIT to exit the Stat Editor.

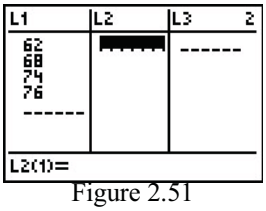



Figure 2.51

b. Obtain the mean, variance, and standard deviation of the pulse rates using the definitions.

Press 2nd LIST ► ► to obtain the LIST MATH menu.

(i) Obtain the mean of the pulse rates.

Select 3: mean( and press ENTER. Press 2nd L1 to select the four pulse rates. Press ENTER. The output from the TI calculator is displayed in Figure 2.53, indicating the mean, 70.

(ii) Obtain the sum of the squared differences.

On the homescreen, press 2nd L1 − 2nd LIST ► ►, selecting 3: mean(. Press 2nd L1 and ). Press STO→ 2nd L2 ENTER, storing the differences between the data value and the mean in list L2. Press 2nd L2, $x^2$. Press STO→ 2nd L3 ENTER, storing the squared differences in list L3. The output from the TI calculator is displayed in Figure 2.52, indicating the sum of the squared difference is 120.

(iii) Obtain the variance.

Press 2nd LIST ► ► to obtain the LIST MATH menu. Select 5: sum( and press ENTER. Press 2nd L3 and ) to obtain the sum of the squared differences in list L3. To obtain the variance, divide this sum by $n - 1$. Do this by pressing 2nd LIST ► ► to obtain the LIST MATH menu. Select 5: sum( and press ENTER. Press 2nd L3 and ). Press ÷, and 2nd LIST ► to obtain the LIST OPS menu. Select 3: dim( and press ENTER. Press 2nd L1 and ). Press ENTER. The output from the TI calculator is displayed in Figure 2.53, indicating the variance, 40.

(iv) Obtain the standard deviation.

Press 2nd $\sqrt{}$ 40 ) to obtain the standard deviation, 6.32, as shown

in Figure 2.54.

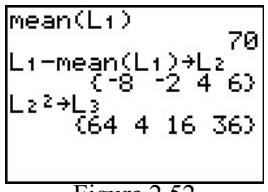


Figure 2.52

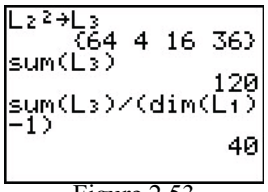


Figure 2.53

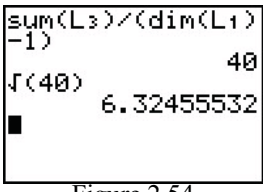


Figure 2.54

Observe that the mean of the pulse rates is 70, the variance is 40, and the standard deviation is 6.32.

3. Obtain the mean, variance, and standard deviation of the pulse rates using the STAT CALC menu..

Press STAT ► to obtain the STAT CALC menu.

a. Select 1: 1-Var Stats and press ENTER. Press 2nd L1 to select the four pulse rates. Press ENTER. The output from the TI calculator is displayed in Figure 2.55.

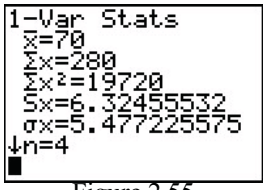


Figure 2.55

Observe that the mean of the pulse rates is 70 and the standard deviation is 6.32.

# Chapter 3
# Relationships Between Quantitative Variables

## Introduction

In this chapter, we will learn how to describe the relationship between two quantitative variables. Remember (from Chapter 2) that the terms quantitative variable and measurement variable are synonyms for data that can be recorded as numerical values and then ordered according to those values. The relationship between weight and height is an example of a relationship between two quantitative variables.
The questions we ask about the relationship between two variables often concern specific numerical features of the association. For example, we may want to know how much weight will increase on average for each 1-inch increase in height. Or, we may want to estimate what the college grade point average will be for a student whose high school grade point average was 3.5.In this chapter, you will learn how to create simple summaries and pictures from various kinds of raw data.

After reading this chapter you should be able to:

1. Display a scatterplot of two quantitative variables.
2. Display subgroups of two quantitative variables on a scatterplot.
3. Display a scatterplot with the regression equation superimposed upon the scatterplot.
4. Make predictions using a regression equation.
5. Obtain the residuals.
6. Find the correlation coefficient and the coefficient of determination for two quantitative variables.
7. Obtain the regression output, identifying the slope, intercept, $r^2$, SSTO, and SSE for two quantitative variables.

**Keystrokes Introduced**

1. [2nd] [STAT PLOT] > scatterplot displays a scatterplot of two quantitative variables.
2. [STAT] CALC> 8: LinReg (a + bx) calculates a regression equation for two quantitative variables.
3. [2nd] [CATALOG] >DiagnosticOn displays $r$, the correlation coefficient, and $r^2$, the coefficient of determination when a linear regression equation is ob-

tained.

4. VARS> 5: Statistics ► ► accesses the regression equation storage registers.
5. STAT >CALC> 1: 1-VarStats analyzes data for one quantitative variable.
6. 2nd LIST >MATH>sum( returns the sum of the elements within a list.

## 3.1 Looking for Patterns With Scatterplots

A scatterplot is a two-dimensional graph of the measurements for two numerical variables. A point on the graph represents the combination of measurements for an individual observation. The vertical axis, which is called the y axis, is used to locate the value of one of the variables. The horizontal axis, called the x axis, is used to locate the value of the other variable.

**Questions to AskAbout a Scatterplot**

What is the average pattern? Does it look like a straight line or is it curved?
What is the direction of the pattern?
How much do individual points vary from the average pattern?
Are there any unusual data points?

**Example 3.1 Height and Handspan**

Tables 3.1a and 3.1b display the observations of a dataset that includes the heights (in inches) and fully stretched hands pans (in centimeters) of 167 college students. The data values for all 167 students are the raw data for studying the connection between height and handspan. Imagine how difficult it is to see the pattern in the data from all 167 observations were shown in Table 3.1. Even when we just look at the data for the first12 students, it takes a while to confirm that there does seem to be a tendency for taller people to have larger handspans.

**Follow these steps** to display a scatterplot of handspan and height measurements for all 167 students.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press Y= and press CLEAR to remove all functions For each line that is not blank, place the cursor on the function and press CLEAR Press 2nd QUIT.

   b. Clear all lists in the Stat editor.

      Press STAT, selecting 4: ClrList. Enter each list name: L1, L2, L3, L4,

L5, L6, as shown in Figure 3.1. Press [ENTER] to execute the command.

| Ss | Height | Hand | Ss | Height | Hand | Ss | Height | Hand |
|---|---|---|---|---|---|---|---|---|
| 1 | 68 | 21.5 | 31 | 67 | 20.0 | 61 | 64 | 20.0 |
| 2 | 71 | 23.5 | 32 | 67 | 20.0 | 62 | 65 | 20.0 |
| 3 | 73 | 22.5 | 33 | 66 | 19.0 | 63 | 74 | 24.0 |
| 4 | 64 | 18.0 | 34 | 62 | 17.0 | 64 | 68 | 21.0 |
| 5 | 68 | 23.5 | 35 | 72 | 22.0 | 65 | 68 | 21.5 |
| 6 | 59 | 20.0 | 36 | 71 | 22.0 | 66 | 69 | 18.5 |
| 7 | 73 | 23.0 | 37 | 61 | 17.5 | 67 | 68 | 23.0 |
| 8 | 75 | 24.5 | 38 | 63 | 19.0 | 68 | 67 | 23.0 |
| 9 | 65 | 21.0 | 39 | 66 | 19.0 | 69 | 61.5 | 20.5 |
| 10 | 69 | 20.5 | 40 | 71 | 22.0 | 70 | 63 | 16.5 |
| 11 | 69 | 20.5 | 41 | 71 | 22.0 | 71 | 67 | 19.5 |
| 12 | 64 | 18.5 | 42 | 66 | 18.5 | 72 | 71 | 23.0 |
| 13 | 67 | 21.0 | 43 | 70 | 20.0 | 73 | 73 | 22.5 |
| 14 | 67 | 19.5 | 44 | 67 | 20.5 | 74 | 63 | 18.5 |
| 15 | 69 | 22.0 | 45 | 69 | 21.0 | 75 | 61 | 18.5 |
| 16 | 73 | 22.0 | 46 | 67 | 19.5 | 76 | 67 | 21.5 |
| 17 | 62 | 20.0 | 47 | 68 | 20.0 | 77 | 72 | 20.5 |
| 18 | 69 | 22.5 | 48 | 67 | 21.5 | 78 | 72 | 20.5 |
| 19 | 64 | 18.5 | 49 | 68 | 22.5 | 79 | 68 | 20.0 |
| 20 | 74 | 21.5 | 50 | 71 | 20.0 | 80 | 66 | 21 |
| 21 | 73 | 24.5 | 51 | 70 | 22.5 | 81 | 67 | 21.5 |
| 22 | 66 | 20.5 | 52 | 74 | 24.5 | 82 | 67 | 20.5 |
| 23 | 74 | 24.5 | 53 | 60 | 18.5 | 83 | 72 | 20.5 |
| 24 | 73 | 21.0 | 54 | 65 | 20.0 | 84 | 67.5 | 21.0 |
| 25 | 69 | 21.0 | 55 | 72 | 24.0 | 85 | 63.75 | 21.5 |
| 26 | 64 | 18.5 | 56 | 76 | 23.5 | 86 | 72 | 21.5 |
| 27 | 67 | 18.0 | 57 | 66 | 21.0 | 87 | 69 | 22.5 |
| 28 | 60 | 19.5 | 58 | 64.5 | 19.5 | 88 | 68 | 21.0 |
| 29 | 75 | 20.5 | 59 | 71 | 20.0 | 89 | 71 | 21.0 |
| 30 | 64 | 21.0 | 60 | 69 | 22.5 | 90 | 71 | 22.0 |

Table 3.1a

2. Enter data using the [STAT] list editor.

Press [STAT] [ENTER] to select the [STAT] list editor.

a. Enter the data for the quantitative variables ”height” and ”handspan” in lists L1 and L2.

Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the height data: 68, 71, 73, ... pressing [ENTER] after each entry.
Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the

hand data: 21.5, 23.5, 22.5, ... in L2 pressing ENTER after each entry, as shown in Figure 5.2.

Figure 3.1

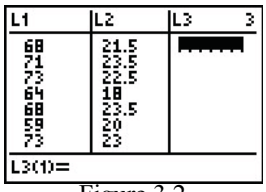


Figure 3.2

| Ss | Height | Hand | Ss | Height | Hand | Ss | Height | Hand |
|---|---|---|---|---|---|---|---|---|
| 91 | 63 | 19.0 | 117 | 67 | 19.5 | 143 | 71 | 18.5 |
| 92 | 70 | 23.0 | 118 | 68 | 22.5 | 144 | 71 | 21.5 |
| 93 | 68 | 20.5 | 119 | 63 | 20.0 | 145 | 63 | 21.0 |
| 94 | 67.5 | 20.5 | 120 | 67 | 21.5 | 146 | 67 | 22.0 |
| 95 | 75 | 21.0 | 121 | 66 | 20.5 | 147 | 65 | 20.5 |
| 96 | 75 | 24.0 | 122 | 72 | 23.5 | 148 | 68 | 19.0 |
| 97 | 71 | 22.0 | 123 | 74 | 22.0 | 149 | 67 | 20.5 |
| 98 | 71 | 21.0 | 124 | 69 | 18.0 | 150 | 73 | 23.0 |
| 99 | 64 | 19.5 | 125 | 68 | 19.0 | 151 | 78 | 25.5 |
| 100 | 71 | 21.0 | 126 | 65 | 19.5 | 152 | 62 | 18.5 |
| 101 | 69 | 19.5 | 127 | 64 | 19.0 | 153 | 70 | 19.0 |
| 102 | 65 | 19.0 | 128 | 67 | 20.0 | 154 | 64 | 19.0 |
| 103 | 69 | 23.0 | 129 | 74 | 23.5 | 155 | 64 | 20.0 |
| 104 | 63 | 20.5 | 130 | 73 | 24.0 | 156 | 72 | 20.5 |
| 105 | 70 | 24.0 | 131 | 64 | 18.5 | 157 | 74 | 24.0 |
| 106 | 71 | 22.0 | 132 | 76 | 24.5 | 158 | 70 | 22.0 |
| 107 | 64 | 20.0 | 133 | 68 | 20.0 | 159 | 70 | 23.5 |
| 108 | 63 | 21.5 | 134 | 76 | 23.0 | 160 | 62 | 17.0 |
| 109 | 65 | 19.0 | 135 | 64.25 | 22.0 | 161 | 64 | 18.5 |
| 110 | 66 | 19.0 | 136 | 69 | 22.5 | 162 | 66 | 20.0 |
| 111 | 66 | 20.0 | 137 | 75 | 24.5 | 163 | 60 | 17.0 |
| 112 | 65 | 19.5 | 138 | 61.5 | 17.0 | 164 | 73 | 23.0 |
| 113 | 67.5 | 20.0 | 139 | 69 | 22.0 | 165 | 66 | 18.5 |
| 114 | 57 | 16.0 | 140 | 67 | 22.0 | 166 | 68 | 21.0 |
| 115 | 72 | 22.5 | 141 | 74 | 24.5 | 167 | 73 | 21.0 |
| 116 | 64 | 17.5 | 142 | 74 | 24.0 | | | |

Table 3.1b

3. Plot the statistical data by creating a scatterplot of handspan and height measurements for all 167 students.

Press 2nd STAT PLOT accessing the stat plot menu.
Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the first row, the scatterplot. Press ENTER. Use the down arrow key to select list

L1 as the list, 2nd L1. Use the down arrow key to enter list L2 as the Ylist: 2nd L2. Use the down arrow key to select the second icon for the mark. The settings for Plot 1 are shown in Figure 3.3.

4. View the graph.

Press ZOOM, 9: ZoomStat to view the graph, as shown in Figure 3.4.

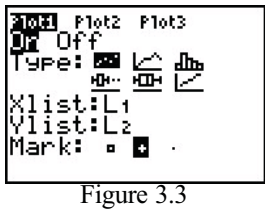


Figure 3.3

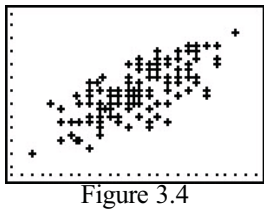


Figure 3.4

5. Save list L1 as HGHT and list L2 as HAND.

Press 2nd L1 STO→ 2nd A-LOCK and type HGHT; press ENTER.
Press 2nd L2 STO→ 2nd A-LOCK and type HAND; press ENTER.

Figure 3.4 is a scatterplot that displays the handspan and height measurements for all 167 students. The hands pan measurements are plotted along the vertical axis (y), and the height measurements are plotted along the horizontal axis (x). Each point represents the two measurements for an individual.

We see that taller people tend to have greater handspan measurements than shorter people do. When two variables tend to increase together, as they do in Figure 3.4, we say that they have a positive association. Another noteworthy characteristic of the graph is that we can describe the general pattern of this relationship with a straight line. In other words, the hands pan and height measurements may have a linear relationship.

## Indicating Groups Within the Data on Scatterplots

When we examined the connection between height and hands pan in Example 3.1, you may have wondered whether we should be concerned about student gender. Both height and hands pan tend to be greater for men than for women, so we should consider the possibility that gender differences might be completely responsible for the observed relationship.

It's easy to indicate subgroups on a scatterplot. We just use different symbols or different colors to represent the different groups.

**Example 3.1 Height and Handspan Continued** The data for females is displayed in Table 3.2. The data for males is displayed in Table 3.3.

Height and HandSpans for 89 Females

| Ss | Height | Hand | Ss | Height | Hand | Ss | Height | Hand |
|---|---|---|---|---|---|---|---|---|
| 1 | 68 | 21.5 | 31 | 64 | 20.0 | 61 | 57 | 16.0 |
| 2 | 64 | 18.0 | 32 | 65 | 20.0 | 62 | 64 | 17.5 |
| 3 | 59 | 20.0 | 33 | 68 | 21.5 | 63 | 67 | 19.5 |
| 4 | 65 | 21.0 | 34 | 69 | 18.5 | 64 | 68 | 22.5 |
| 5 | 69 | 20.5 | 35 | 68 | 23.0 | 65 | 63 | 20.0 |
| 6 | 64 | 18.5 | 36 | 61.5 | 20.5 | 66 | 66 | 20.5 |
| 7 | 67 | 21.0 | 37 | 63 | 16.5 | 67 | 68 | 19.0 |
| 8 | 67 | 19.5 | 38 | 67 | 19.5 | 68 | 65 | 19.5 |
| 9 | 62 | 20.0 | 39 | 63 | 18.5 | 69 | 64 | 19.0 |
| 10 | 64 | 18.5 | 40 | 61 | 18.5 | 70 | 67 | 20.0 |
| 11 | 66 | 20.5 | 41 | 68 | 20.0 | 71 | 64 | 18.5 |
| 12 | 64 | 18.5 | 42 | 66 | 21.0 | 72 | 68 | 20.0 |
| 13 | 67 | 18.0 | 43 | 67 | 20.5 | 73 | 64.25 | 22.0 |
| 14 | 60 | 19.5 | 44 | 63.75 | 21.5 | 74 | 61.5 | 17.0 |
| 15 | 64 | 21.0 | 45 | 72 | 21.5 | 75 | 71 | 18.5 |
| 16 | 67 | 20.0 | 46 | 68 | 21.0 | 76 | 63 | 21.0 |
| 17 | 67 | 20.0 | 47 | 63 | 19.0 | 77 | 65 | 20.5 |
| 18 | 66 | 19.0 | 48 | 68 | 20.5 | 78 | 68 | 19.0 |
| 19 | 62 | 17.0 | 49 | 67.5 | 20.5 | 79 | 67 | 20.5 |
| 20 | 61 | 17.5 | 50 | 64 | 19.5 | 80 | 62 | 18.5 |
| 21 | 63 | 19.0 | 51 | 69 | 19.5 | 81 | 70 | 19.0 |
| 22 | 66 | 18.5 | 52 | 65 | 19.0 | 82 | 64 | 19.0 |
| 23 | 67 | 20.5 | 53 | 63 | 20.5 | 83 | 64 | 20.0 |
| 24 | 67 | 19.5 | 54 | 64 | 20.0 | 84 | 62 | 17.0 |
| 25 | 68 | 20.0 | 55 | 63 | 21.5 | 85 | 64 | 18.5 |
| 26 | 71 | 20.0 | 56 | 65 | 19.0 | 86 | 66 | 20.0 |
| 27 | 60 | 18.5 | 57 | 66 | 19.0 | 87 | 60 | 17.0 |
| 28 | 65 | 20.0 | 58 | 66 | 20.0 | 88 | 66 | 18.5 |
| 29 | 66 | 21.0 | 59 | 65 | 19.5 | 89 | 68 | 21.0 |
| 30 | 64.5 | 19.5 | 60 | 67 | 20.0 | | | |

Table 3.2

**Follow these steps** to display a scatterplot of handspan and height measurements for the 89 female students and the 78 male students.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press Y= and press CLEAR to remove all functions For each line that is not blank, place the cursor on the function and press CLEAR Press 2nd QUIT.

b. Clear all lists in the Stat editor.

Press STAT, selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6, as shown in Figure 3.5. Press ENTER to execute the command.

Height and HandSpans for 78 Males

| Ss | Height | Hand | Ss | Height | Hand | Ss | Height | Hand |
|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 23.5 | 27 | 72 | 24.0 | 53 | 71 | 22.0 |
| 2 | 73 | 22.5 | 28 | 76 | 23.5 | 54 | 72 | 22.5 |
| 3 | 68 | 23.5 | 29 | 71 | 20.0 | 55 | 67 | 21.5 |
| 4 | 73 | 23.0 | 30 | 69 | 22.5 | 56 | 72 | 23.5 |
| 5 | 75 | 24.5 | 31 | 74 | 24.0 | 57 | 74 | 22.0 |
| 6 | 69 | 20.5 | 32 | 68 | 21.0 | 58 | 69 | 18.0 |
| 7 | 69 | 22.0 | 33 | 67 | 23.0 | 59 | 74 | 23.5 |
| 8 | 73 | 22.0 | 34 | 71 | 23.0 | 60 | 73 | 24.0 |
| 9 | 69 | 22.5 | 35 | 73 | 22.5 | 61 | 76 | 24.5 |
| 10 | 74 | 21.5 | 36 | 67 | 21.5 | 62 | 76 | 23.0 |
| 11 | 73 | 24.5 | 37 | 72 | 20.5 | 63 | 69 | 22.5 |
| 12 | 74 | 24.5 | 38 | 72 | 20.5 | 64 | 75 | 24.5 |
| 13 | 73 | 21.0 | 39 | 67 | 21.5 | 65 | 69 | 22.0 |
| 14 | 69 | 21.0 | 40 | 72 | 20.5 | 66 | 67 | 22.0 |
| 15 | 75 | 20.5 | 41 | 67.5 | 21.0 | 67 | 74 | 24.5 |
| 16 | 72 | 22.0 | 42 | 69 | 22.5 | 68 | 74 | 24.0 |
| 17 | 71 | 22.0 | 43 | 71 | 21.0 | 69 | 71 | 21.5 |
| 18 | 66 | 19.0 | 44 | 71 | 22.0 | 70 | 67 | 22.0 |
| 19 | 71 | 22.0 | 45 | 70 | 23.0 | 71 | 73 | 23.0 |
| 20 | 71 | 22.0 | 46 | 75 | 21.0 | 72 | 78 | 25.5 |
| 21 | 70 | 20.0 | 47 | 74 | 24.0 | 73 | 72 | 20.5 |
| 22 | 69 | 21.0 | 48 | 71 | 22.0 | 74 | 74 | 24.0 |
| 23 | 67 | 21.5 | 49 | 71 | 21.0 | 75 | 70 | 22.0 |
| 24 | 68 | 22.5 | 50 | 71 | 21.0 | 76 | 70 | 23.5 |
| 25 | 70 | 22.5 | 51 | 69 | 23.0 | 77 | 73 | 23.0 |
| 26 | 74 | 24.5 | 52 | 70 | 24.0 | 78 | 73 | 21.0 |

Table 3.3

2. Enter data using the STAT list editor.

Press STAT ENTER to select the STAT list editor.

a. Enter the data for the quantitative variables ”height” and ”handspan” for females in lists L1 and L2.

Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the height data for females: 68, 64, 59, ... pressing ENTER after each entry.

Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the hand data for females: 21.5, 18.0, 20.0, ... in L2 pressing ENTER after each entry, as shown in Figure 3.6.

b. Enter the data for the quantitative variables "height" and "handspan" for males in lists L3 and L4.

Place the cursor on list L3 row 1 to make L3(1) the active list row. Enter the height data for males: 71, 73, 68, ... pressing ENTER after each entry. Place the cursor on list L4 row 1 to make L4(1) the active list row. Enter the hand data for males: 23.5, 22.5, 23.5, ... in L2 pressing ENTER after each entry, as shown in Figure 5.7.

Figure 3.5

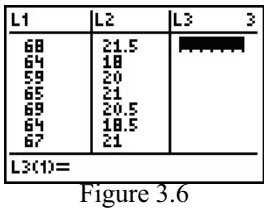


Figure 3.6

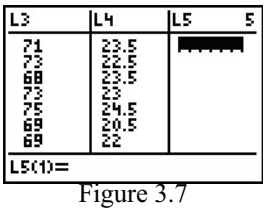


Figure 3.7

3. Plot the statistical data by creating a scatterplot indicating groups within the data.

Press 2nd STAT PLOT accessing the StatPlot menu.

Create a scatterplot of female heights and handspans with heights on the horizontal axis and handspan on the vertical axis.
Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the first row, the scatterplot. Press ENTER. Use the down arrow key to select list L1 as the Xlist, 2nd L1. Use the down arrow key to enter list L2 as the Ylist: 2nd L2. Use the down arrow key to select the second icon for the mark. The settings for Plot 1 are shown in Figure 3.8.

Create a scatterplot of male heights and handspans with heights on the horizontal axis and handspan on the vertical axis.
Use the up arrow key, to select Plot 2. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the first row, the scatterplot. Press ENTER. Use the down arrow key to select list L3 as the Xlist, 2nd L3. Use the down arrow key to enter list L4 as the Ylist: 2nd L2. Use the down arrow key to select the third icon for the mark. The settings for Plot 2 are shown in Figure 3.9.

4. Set the Window viewing variables in order to view the graph.

Press [WINDOW], row 1, column 2. Set Xmin to 55. Set Xmax to 80; Xscl to 1; Ymin to 15. Set Ymax to 26; Yscl to 10; Xres to 1. These settings are illustrated in Figure 3.10

5. View the graph.

Press [GRAPH] to view the graph, as shown in Figure 3.11.

Figure 3.8

Figure 3.9

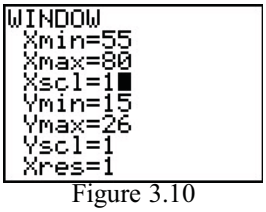


Figure 3.10

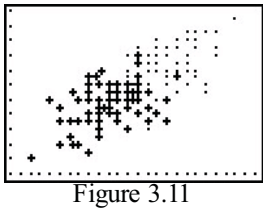


Figure 3.11

6. Save list L1 as HGHTF and list L2 as HANDF.

Press [2nd] [L1] [STO→] [2nd] [A-LOCK] and type HGHTF; press [ENTER].
Press [2nd] [L2] [STO→] [2nd] [A-LOCK] and type HANDF; press [ENTER].

7. Save list L3 as HGHTM and list L4 as HANDM.

Press [2nd] [L1] [STO→] [2nd] [A-LOCK] and type HGHTM; press [ENTER].
Press [2nd] [L2] [STO→] [2nd] [A-LOCK] and type HANDM; press [ENTER].

Notice that the positive association between hands pan and height appears to hold within each sex. For both men and women, hands pan tends to increase as height increases.

## 3.2 Describing Linear Patterns With a Regression Line

Scatter plots show us a lot about a relationship, but we often want more specific numerical descriptions of how the response and explanatory variables are related. Imagine, for example, that we are examining the weights and heights of a sample of college women. We might want to know what the increase in average weight is for each I-inch increase in height. Or, we might want to estimate the average weight for women with a specific height, like 5'10".

Regression analysis is the area of statistics used to examine the relationship between a quantitative response variable and one or more explanatory variables. A key element of regression analysis is the *estimation* of a regression equation that describes how, on average, the response variable is related to the explanatory vari-

ables. This regression equation can be used to answer the types of questions that we just asked about the weights and heights of college women.

A regression equation can atso-be used to predict values of a response variable using known values of an explanatory variable. For instance, it might be useful for colleges to have an equation for the connection between verbal SAT score and college grade point average (GPA). They could use that equation to predict the potential GPAs of future students, based on their verbal SAT scores. Some colleges actually do this kind of prediction to decide whom to admit, but they use a collection of variables to predict GPA.

There are many types of relationships and many types of regression equations. The simplest kind of relationship between two variables is a straight line, and that's the only type we will discuss here. Straight-line relationships occur frequently in practice, so this is a useful and important type of regression equation. Before we use a straight-line regression model, however, we should always examine a scatterplot to verify that the pattern actually is linear.

**Example 3.2 Driver Age and the Maximum Legibility Distance of Highway Signs** In a study of the legibility and visibility of highway signs, a Pennsylvania research firm determined the maximum distance at which each of 30 drivers could read a newly designed sign. The 30 participants in the study ranged in age from 18 to 82 years old. The government agency that funded the research hoped to improve highway safety for older drivers and wanted to examine the relationship between age and the sign legibility distance.

Table 5.4 lists the data. We will use the TI calculator to display a scatterplot to show that the relationship between "maximum distance" and "age" has a straight line pattern and to find the "best" line for this set of measurements. We will display a line that describes the average relationship between the two variables.

| Age | Distance | Age | Distance | Age | Distance |
|---|---|---|---|---|---|
| 18 | 510 | 37 | 420 | 68 | 300 |
| 20 | 590 | 41 | 460 | 70 | 390 |
| 22 | 560 | 46 | 450 | 71 | 320 |
| 23 | 510 | 49 | 380 | 72 | 370 |
| 23 | 460 | 53 | 460 | 73 | 280 |
| 25 | 490 | 55 | 420 | 74 | 420 |
| 27 | 560 | 63 | 350 | 75 | 460 |
| 28 | 510 | 65 | 420 | 77 | 360 |
| 29 | 460 | 66 | 300 | 79 | 310 |
| 32 | 410 | 67 | 410 | 82 | 360 |

Table 3.4

**Follow these steps** to display a scatterplot with the regression equation superimposed upon the scatterplot.

1. Preparations:

a. Turn off all"Y=" functions.

Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

b. Clear all lists in the Stat editor.

Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

c. Turn Diagnostics On to display $r$, the correlation coefficient, and $r^2$, the coefficient of determination.

Press [2nd] [CATALOG], located on the bottom row, 2nd column from the left above the [0]. Press [ALPHA] [D], and use the down arrow key to locate **DiagnosticOn**, as shown in Figure 5.12. Press [ENTER] to select the command and press [ENTER] once again to execute the command.

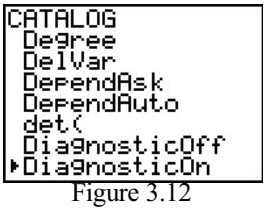


Figure 3.12

2. Enter data using the [STAT] list editor.

Press [STAT] [ENTER] to select the [STAT] list editor.

a. Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the "Age" data: 18, 20, 22, ... pressing [ENTER] after each entry.
Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the "Distance" data: 510, 590, 560, ... in L2 pressing [ENTER] after each entry, as shown in Figure 5.13.

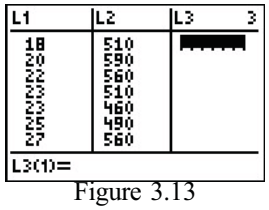


Figure 3.13

3. Obtain the regression equation.

Press [STAT] [▶] to obtain the [STAT] CALC menu.

a. Use the down arrow key, ▼, seven times and press ENTER, or just press 8 to select 8: LinReg (a+bx), as shown in Figure 3.14. Press 2nd L1 to select the "Age" data. Press , 2nd L2 to select the "Distance" data, as shown in Figure 5.15. Press ENTER to execute the command. The output from the TI calculator is displayed in Figure 5.16.

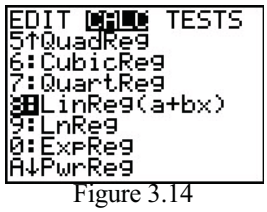


Figure 3.14

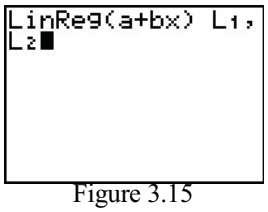


Figure 3.15

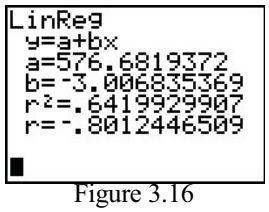


Figure 3.16

The regression line y=577 - 3x describes how the maximum sign legibility distance (the y variable) is related to driver age (the x variable).

4. Obtain data points to plot the regression equation.

Press VARS, row 4, column 4. Select 5: Statistics. Use the right arrow, ►, twice, selecting 2: a. Press +, VARS, 5: Statistics, and the right arrow, ►, twice, selecting 3: b. Press ×, 2nd L1. Press STO→ 2nd L3. Your screen should look like Figure 5.17. These data points represent the predicted values of "Distance" from the "Age" variable stored in list L1. These predicted values of "Distance" are stored in L3.

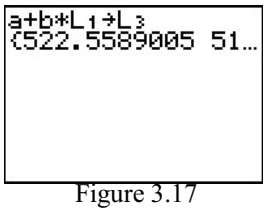


Figure 3.17

5. Display a scatterplot with the regression equation superimposed upon the scatterplot.

Press 2nd STAT PLOT accessing the StatPlot menu.

Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the first row, the scatterplot. Press ENTER. Use the down arrow key to select list L1 as the Xlist, 2nd L1. Use the down arrow key to enter list L2 as the Ylist: 2nd L2. Use the down arrow key to select the second icon for the mark. The settings for Plot 1 are shown in Figure 3.18.

Use the up arrow key to place the cursor on Plot2. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the second icon in the first row, the xyLine. Press ENTER. Use the down arrow key to select L1 as the Xlis t, 2nd L1. Use the down arrow key to select L3 as the Ylist:, 2nd L3. The settings for Plot 2 are shown in Figure 3.19.

6. View the graph.

Press ZOOM, 9: ZoomStat to view the graph, as shown in Figure 3.20.

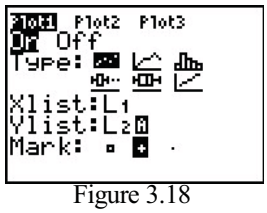


Figure 3.18

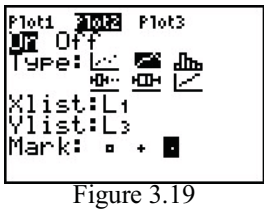


Figure 3.19

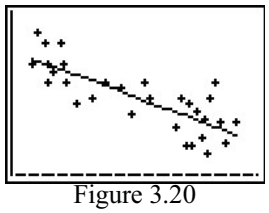


Figure 3.20

Earlier, we asked these two questions about distance and age:

(i) How much does the distance decrease when age is increased?

(ii) For drivers of any specific age, what is the average distance at which the sign can be read?

The slope of the equation can be used to answer the first question. Remember that the slope is the number that multiplies the x varia,ble and the sign of the slope indicates the direction of the association. Here, the slope tells us that, on average, the legibility distance decreases 3 feet when age increases by one year. This information can be used to estimate the average change in distance for any difference in ages, For an age increase of 30 years, the estimated decrease in legibility distance is 90 feet because the slope is -3 feet per year.

The question about estimating the average legibility distances for a specific age is answered by using the specific age as the x value in the regression equation. To emphasize this use of the regression line, we write it as

$$Average\ distance = 577 - 3\ Age$$

1. Make predictions for specific ages, 20, 50, and 80, finding the average distance at which the sign can be read.

Press VARS, row 4, column 4. Select 5: Statistics. Use the right arrow, ▶, twice, selecting 2: a. Press +, VARS, 5: Statistics, and the right arrow, ▶, twice, selecting 3: b. Press ×, 20. Press ENTER.
Press VARS, row 4, column 4. Select 5: Statistics. Use the right arrow, ▶, twice, selecting 2: a. Press +, VARS, 5: Statistics, and the right arrow, ▶, twice, selecting 3: b. Press ×, 50. Press ENTER.
Press VARS, row 4, column 4. Select 5: Statistics. Use the right arrow, ▶, twice, selecting 2: a. Press +, VARS, 5: Statistics, and the right arrow, ▶, twice, selecting 3: b. Press ×, 80. Press ENTER. The results of these three

calculations are shown in Figure 3.21.

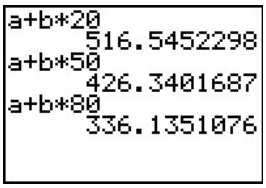



Figure 3.21

For any given line, we can calculate the predicted value $\widehat{y}$ for each point in the observed data. To do this for any particular point, we use the observed $x$ value in the regression equation. The prediction error for an observation is the difference between the observed $y$ value and the predicted value $\widehat{y}$; the formula is $error = (y - \widehat{y})$. The terminology "error" is somewhat misleading, since the amount by which an individual differs from the line is usually due to natural variation rather than "errors" in the measurements. A more neutral term for the difference $(y - \widehat{y})$ is that it is the residual for that individual.

2. Obtain the residuals.

Recall that the predicted values of "Distance", based upon the "Age" variable are stored in L3 and the observed values of "Distance" are in L2.

Press STAT ENTER to select the STAT list editor.

a. Place the cursor at the top of list L4. Press 2nd L2, - 2nd L3, as shown in Figure 3.22, pressing ENTER to obtain the residuals.
b. Place the cursor at the top of list L5. Press 2nd LIST, selecting the list RESID, as shown in Figure 3.23. Press ENTER to drive the residuals, that are automatically generated on the TI calculator with that list name, into L5. Observe that the residuals displayed in lists L4 and L5 are identical, as shown in Figure 3.24.

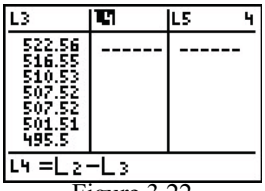



Figure 3.22

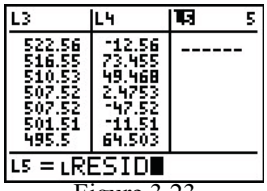



Figure 3.23

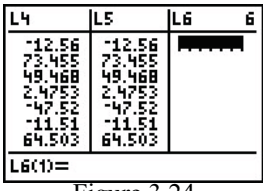



Figure 3.24

3. Turn off all plots and return the graph window to standard viewing.

Press 2nd STAT PLOT, selecting PlotsOff and press ENTER. Press ZOOM and select 6: ZStandard to restore the default graph window settings.

## 3.3 Measuring Strength and Direction with Correlation

The linear pattern is so common that a statistic was created to characterize this type of relationship. The statistical correlation between two quantitative variables is a number that *indicates the strength and the direction of a straight-line relationship*.

(i) The *strength* of the relationship is determined by the *closeness of the points to a straight line*.

(ii) The *direction* is determined by whether one variable generally increases or generally decreases when the other variable increases.

As used in statistics, the meaning of the word correlation is much more specific than it is in everyday life. A statistical correlation only describes linear relationships. Whenever a correlation is calculated, a straight line is used as the frame of reference for evaluating the relationship. When

**Example 3.2 Driver Age and the Maximum Legibility Distance of Highway Signs Revisted** In a study of the legibility and visibility of highway signs, a Pennsylvania research firm determined the maximum distance at which each of 30 drivers could read a newly designed sign. The 30 participants in the study ranged in age from 18 to 82 years old. The government agency that funded the research hoped to improve highway safety for older drivers and wanted to examine the relationship between age and the sign legibility distance.
Table 3.3 lists the data. We will use the TI calculator to determine the correlation coefficient between "maximum distance" and "age".

**Follow these steps** to find the correlation coefficient and the coefficient of determination.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor: Caution: If the "Age" and "Distance" data are within L1 and L2, **Do NOT** execute this step .

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.
   c. Turn Diagnostics On to display $r$, the correlation coefficient, and $r^2$, the coefficient of determination.

      Press [2nd] [CATALOG], located on the bottom row, 2nd column from the left above the 0. Press [ALPHA] [D], and use the down arrow key to

locate **DiagnosticOn**, as shown in Figure 3.12. Press ENTER to select the command and press ENTER once again to execute the command.

2. Enter data using the STAT list editor.

   Press STAT ENTER to select the STAT list editor.

   a. Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the "Age" data: 18, 20, 22, ... pressing ENTER after each entry.
   Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the "Distance" data: 510, 590, 560, ... in L2 pressing ENTER after each entry, as shown in Figure 3.13.

3. Obtain the regression equation, correlation coefficient and the coefficient of determination.

   Press STAT ► to obtain the STAT CALC menu.

   a. Use the down arrow key, ▼, seven times and press ENTER, or just press 8 to select 8: LinReg (a+bx), as shown in Figure 3.14. Press 2nd L1 to select the "Age" data. Press , 2nd L2 to select the "Distance" data, as shown in Figure 3.15. Press ENTER to execute the command. The output from the TI calculator is displayed in Figure 3.16 and Figure 3.25.

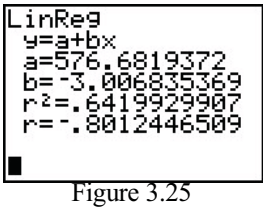



Figure 3.25

For the data shown in Figure 3.20 relating driver age and sign legibility distance, the correlation is $r = -0.80$. This value indicates a somewhat strong negative association between the variables.

## Calculating the Sum of Squared Errors

A least squares line has the property that the sum of squared differences between the observed values of y and the predicted values is smaller for that line than it is for any other line. Put more simply, the least squares line minimizes the sum of squared prediction errors for the observed data set. The notation SSE, which stands for sum of squared errors, is used to represent the sum of squared prediction errors. The least squares line (the regression line) has a smaller SSE than any other regression line that might be used to predict the response variable.

**Example - Exam Scores** Suppose that $x$ = score on exam 1 in a course and $y$= score on exam 2, and that the first two rows in Table 5.5 (shown below) give x values and y values for n = 6 students. We will use the TI calculator to obtain the regression output, identifying the slope, intercept, $r^2$, SSTO, and SSE for this set of measurements.

| | | | | | | |
|---|---|---|---|---|---|---|
| x = Exam 1 score | 70 | 75 | 80 | 80 | 85 | 90 |
| y = Exam 2 score | 75 | 82 | 80 | 86 | 90 | 91 |

Table 3.5

**Follow these steps** to obtain the regression output, identifying the slope, intercept, $r^2$, SSTO, and SSE for this set of measurements.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

   c. Turn Diagnostics On to display $r$, the correlation coefficient, and $r^2$, the coefficient of determination.

   Press [2nd] [CATALOG], located on the bottom row, 2nd column from the left above the [0]. Press [ALPHA] [D], and use the down arrow key to locate **DiagnosticOn**, as shown in Figure 5.12. Press [ENTER] to select the command and press [ENTER] once again to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the "$x$ = Exam 1 score" data: 70, 75, 80, ... pressing [ENTER] after each entry.
   Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the "$y$ =Exam 2 score" data: 75, 82, 80, ... in L2 pressing [ENTER] after

each entry, as shown in Figure 3.13.

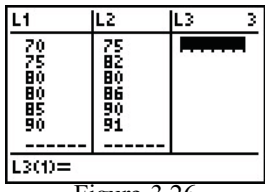


Figure 3.26

3. Obtain the regression equation.

Press STAT ► to obtain the STAT CALC menu.

a. Use the down arrow key, ▼, seven times and press ENTER, or just press 8 to select 8: LinReg (a+bx), as shown in Figure 3.27. Press 2nd L1 to select the "$x$ = Exam 1 score" data. Press , 2nd L2 to select the "$y$ =Exam 2 score" data, as shown in Figure 3.28. Press ENTER to execute the command. The output from the TI calculator is displayed in Figure .29.

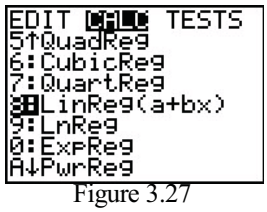


Figure 3.27

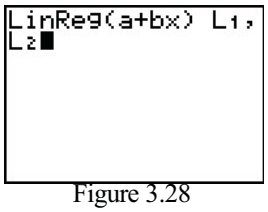


Figure 3.28

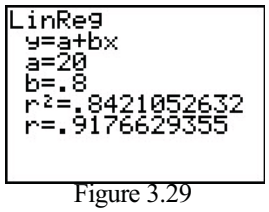


Figure 3.29

The regression equation is y = 20 + 0.8x; the $y$-intercept is 20 and the slope is 0.8. The correlation coefficient, $r = .918$ describes a moderately strong positive association. The squared correlation is $r = (.918)^2 = 0.842$. The "$x$ = Exam 1 score" explains 84.2% of the variation among the "$y$ =Exam 2 score" data.

4. Obtain the sum of square errors.

Press VARS, row 4, column 4. Select 5: Statistics. Use the right arrow, ►, twice, selecting 2: a. Press ⊞, VARS, 5: Statistics, and the right arrow, ►, twice, selecting 3: b. Press ×, 2nd L1. Press STO→ 2nd L3. Your screen should look like Figure 5.30. These data points represent the predicted values, $\widehat{y}$, from "$x$ =Exam 1 score" variable stored in list L1. These predicted values, $\widehat{y}$, of "Exam 2 score" are stored in L3.

Press STAT ENTER to select the STAT list editor.

a. Place the cursor at the top of list L4. Press 2nd L2 , − 2nd L3 , as shown in Figure 3.31, pressing ENTER to obtain the residuals. The resid-

uals are shown in Figure 3.32.

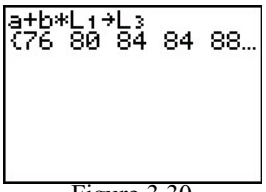


Figure 3.30

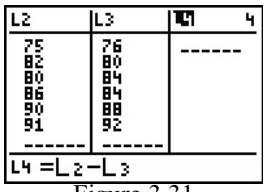


Figure 3.31

b. Press 2nd QUIT. Press 2nd LIST ▶ ▶, selecting 5: sum(. Press 2nd L4 $x^2$ ). Press ENTER . Your screen should look like Figure 3.33.

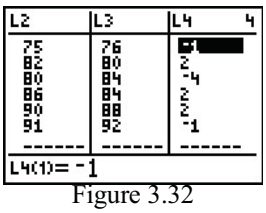


Figure 3.32

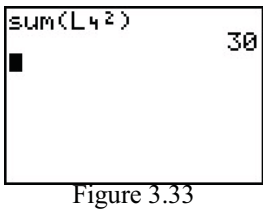


Figure 3.33

The sum of squared errors is SSE = 30.

5. Obtain the total sum of squares, $SSTO = sum\ ((y\text{-}\overline{y})^2)$.

To obtain the mean of the predicted $y$ values, press STAT > CALC, selecting 1: 1Var Stats. Press ENTER. Press 2nd L3 and ENTER, as shown in Figure 5.34. The results are shown in Figure 5.35.

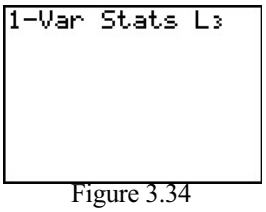


Figure 3.34

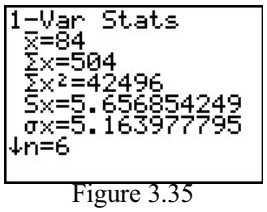


Figure 3.35

The mean of the predicted $y$ values is 84.

Press STAT ENTER to select the STAT list editor.

a. Place the cursor at the top of list L5. Press 2nd L2 , - 8 4 , as shown in Figure 3.36, pressing ENTER. The results are shown in Figure 3.37.

b. Press 2nd QUIT. Press 2nd LIST ▶ ▶, selecting 5: sum(. Press 2nd L5 $x^2$ ). Press ENTER. Your screen should look like Figure

3.38.

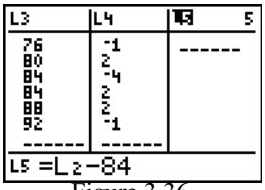


Figure 3.36

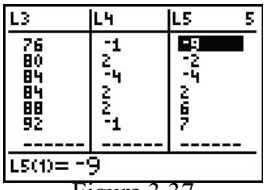


Figure 3.37

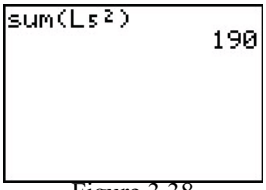


Figure 3.38

The total sum of quares, $SSTO = sum\left((y\text{-}\overline{y})^2\right)$ is 190.

The coefficient of determination, $r^2 = \frac{SSTO-SSE}{SSTO} = \frac{190-30}{190} = 0.84211$.

# Chapter 4
# Relationships Between Categorical Variables

This chapter is about the analysis of the relationship between two categorical variables, so let's begin by recalling the meaning of the term *categorical variable*.The raw data from categorical variables consist of group or category names that don't necessarily have any ordering. Eye color and hair color, for instance, are categorical variables.

We can also use the methods of this chapter to examine *ordinal* variables. Ordinal variables can be thought of as categorical variables for which the categories have a natural ordering. For example, a researcher might define categories for quantitative variables, like age, income, or years of education.

Although there are many questions that we can and will ask about two categorical variables, in most cases the principal question that we ask is: Is there a relationship between the two variables, so that the category into which individuals fall for one variable seems to depend on the category they are in for the other variable?

After reading this chapter you should be able to:

1. Construct a table of Frequency Counts from raw data including row and column percents.
2. Conduct a chi-square test, including finding observed counts, computing a chi-square statistic and find the p-value.
3. Find a $p$-value given the chi-square value and degrees of freedom.

**Keystrokes Introduced**

1. [2nd] [LIST] [►] OPS>3: dim(*listname*) returns the dimension (number of elements) of listname.
2. [2nd] [LIST] [►] [►] MATH > 5: sum(*list* [,*start*,*end* ]) returns the sum of the elements of *list* from *start* to *end*.
3. [2nd] [MATRIX] [►] [►], selecting the [MATRIX] EDIT menu. This command enables you to edit a matrix element value including the dimensions and the elements of the matrix.
4. [2nd] [MATRIX] >NAMES selecting a matrix, pressing [ENTER] to view the elements of the matrix.
5. [STAT] [►] [►], selecting the [STAT] TEST menu. You will select C: $\chi^2$ Test, performing a $\chi^2$ test, where the observations have been entered into matrix A.
6. [2nd] [DISTR], using the down arrow key, [▼], several times to select 7: $\chi^2$cdf(. The arguments in $\chi^2$cdf are(*lowerbound*,*upperbound*,*df* ). The command com-

putes the distribution probability between *lowerbound* and *upperbound* for the specified degrees of freedom *df*.

## 4.1 Displaying Relationships

We have already encountered several examples of the type of problem we will study in this chapter. In Chapter 2, for instance, we described a study of 479 children that found that children who slept either with a nightlight or in a fullylit room before the age of two had a higher incidence of myopia (nearsightedness) later in childhood. Data from the PennState1 worksheet will be used to illustrate how relationships between categorical variables may be presented.

**Example**

In an experiment done in a statistics class, 92 college students were given a form read "Randomly choose one of the letters S or Q." Another 98 students were given a form with the order ofthe letters reversed, to read "Randomly choose one of the letters Q or S." The purpose was to determine if the order of listing the letters might influence the choice of letters. The possible influence of the order of listing items is a concern in elections. Many election analysts feel a candidate gains an advantage if he or she is the first candidate listed on the ballot. The data is contained in the *PennState1* worksheet and is displayed in Table 4.1 and Table 4.2.

S S S S S S S S S S S S S S S S S S S
S S S S S S S S S S S S S S S S S S S
S S S S S S S S S S S S S S S S S S S
S S S S Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q
Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q

Table 4.1 "Randomly pick a letter-S or Q"

S S S S S S S S S S S S S S S S S S S
S S S S S S S S S S S S S S S S S S S
S S S S S S S Q Q Q Q Q Q Q Q Q Q Q Q
Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q
Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q
Q Q Q

Table 4.2 "Randomly pick a letter-Q or S"

TI calculators allow only for numerical values to be used in a statistical analysis. We **can not** use the letters "S" or "Q" since these letters are replaced by the value stored in memory for the "S" and "Q" variables in the calculator.
The solution to the problem is to assign a unique numerical code for each value of the variable. In this case, you might code "S = 0" and "Q = 1" on the TI calculator.

**Follow these steps** to construct a table of frequency counts from the raw data, including row and column percents.

Chapter 4 Relationships Between Categorical Variables

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the data found in Table 4.1 "Randomly pick a letter-S or Q".

      Place the cursor on list L1 row 1 to make L1(1) the active list row. Code "S = 0" and "Q = 1" entering 0 for "S" and 1 for "Q" pressing [ENTER] after each entry.
   b. Enter the data found in Table 4.2 "Randomly pick a letter-Q or S".

      Place the cursor on list L2 row 1 to make L2(1) the active list row. Code "S = 0" and "Q = 1" entering 0 for "S" and 1 for "Q" pressing [ENTER] after each entry, as shown in Figure 4.1.

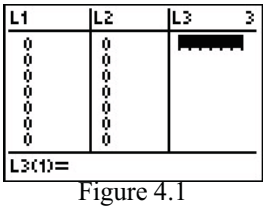


Figure 4.1

3. Count the number of observations entered into list L1 and list L2.

   On the homescreen press [2nd] [LIST] [▶] to select the [LIST] OPS menu. Select 3: dim(. Press [ENTER]. Press [2nd] [L1] [)]. Press [ENTER] to execute the command.We know that there are 92 entries in list L1, corresponding to the 92 students, as shown in Figure 4.2.
   On the homescreen press [2nd] [LIST] [▶] to select the [LIST] OPS menu. Select 3: dim(. Press [ENTER]. Press [2nd] [L2] [)]. Press [ENTER] to execute the command.We know that there are 98 entries in list L2, corresponding to the

98 students, as shown in Figure 4.2.

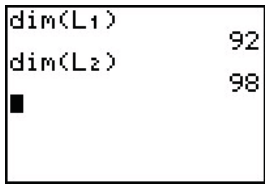



Figure 4.2

4. Count the number of 0's or the number of occurrences of the letter' 'Q".

On the homescreen press 2nd LIST ► ► to select the LIST MATH menu. Select 5: sum(. Press ENTER. Press 2nd L1 ). Press ENTER to execute the command. Since the letter "Q" was coded as a "1", we know that "Q" occurred 31 times out of the 92 responses, as shown in Figure 4.3.
On the homescreen press 2nd LIST ► ► to select the LIST MATH menu. Select 5: sum(. Press ENTER. Press 2nd L2 ). Press ENTER to execute the command. Since the letter "Q" was coded as a "1", we know that "Q" occurred 53 times out of the 98 responses, as shown in Figure 4.3.

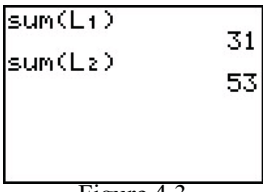



Figure 4.3

As a result of counting the number of occurrences of the letter "Q", you are now able to construct a frequency table of the number of occurrences of the letter "Q" and the letter "S". This information is shown in Table 4.3.

| | Letters picked | | |
|---|---|---|---|
| Form | S | Q | Total |
| S first | 61 | 31 | 92 |
| Q first | 45 | 53 | 98 |
| Total | 106 | 84 | 190 |

Table 4.3 Occurrences of the letters "S" and "Q"

**Follow these steps** to construct a table of row percents based upon Table 4.3.

1. Enter the data found in Table 4.3 using the STAT list editor.

Press STAT ENTER to select the STAT list editor.

a. Place the cursor on list L3 row 1 to make L3(1) the active list row. Focus on the row labeled "S first." Enter the count for "S",61 and the count for "Q", 31, pressing [ENTER] after each entry.

b. Place the cursor on list L5 row 1 to make L5(1) the active list row. Focus on the row labeled "Q first." Enter the count for "S",45 and the count for "Q", 53, pressing [ENTER] after each entry.

2. Calculate the row percents for "S first".

On the homescreen, press [2nd] [L3] [÷] [2nd] [LIST] [►] [►] selecting the [LIST] MATH menu. Select 5: sum( and press [2nd] [L3] [)] [×] 100 [STO→] [2nd] [L4], as shown in Figure 4.4.

3. Calculate the row percents for "Q first."

On the homescreen, press [2nd] [L5] [÷] [2nd] [LIST] [►] [►] selecting the [LIST] MATH menu. Select 5: sum( and press [2nd] [L5] [)] [×] 100 [STO→] [2nd] [L6], as shown in Figure 4.5.

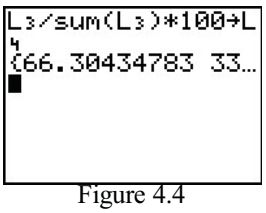


Figure 4.4

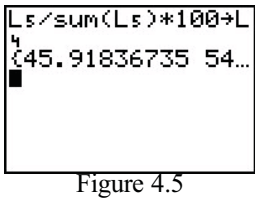


Figure 4.5

4. Select [STAT] [ENTER] to select the [STAT] list editor.

a. Place the cursor on list L4 row 1 to make L4(1) the active list row. View the row percents for S first displayed in list L4, as shown in Figure 4.6.

b. Place the cursor on list L6 row 1 to make L6(1) the active list row. View the row percents for Q first displayed in list L6, as shown in Figure 4.6.

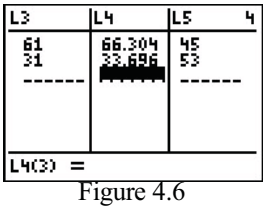


Figure 4.6

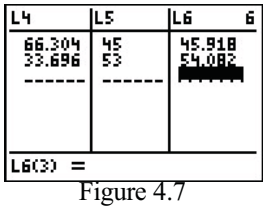


Figure 4.7

As a result of the calculation of the row percents, you are now able to add the row percent to the table of the number of occurrences of the letter "Q" and the letter "S". This information is shown in Table 4.4. The % of Row

for Total were calculated seperately.

| | Letters picked | | |
|---|---|---|---|
| Form | S | Q | Total |
| S first | 61 | 31 | 92 |
| % of Row | 66.3 | 33.7 | 100.0 |
| | | | |
| Q first | 45 | 53 | 98 |
| % of Row | 45.9 | 54.1% | 100.0 |
| | | | |
| Total | 106 | 84 | 190 |
| | 55.8 | 44.2 | 100.0 |

Table 4.4 Occurrences of the letters "S" and "Q"

We can use row percents to compare the rates of the letters picked by those who received the form "S first" and "Q first." The first row of the table gives the data for those who received the form "S first." Among the 92 individuals who received the form "S first", 66.3% picked the letter S, and 33.7% picked the letter Q. The second row of the table gives the data for those who received the form "Q first." Among the 98 individuals who received the form "Q first", 45.9% picked the letter S, and 54.1% picked the letter Q. The difference between the two sets of row percents appears to indicate a relationship. There is a relationship between two categorical variables forming a two-way table if two or more rows have different distributions of row percents.

**Follow these steps** to construct a table of column percents based upon Table 4.3.

1. Enter the data found in Table 4.3 using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Place the cursor on list L3 row 1 to make L3(1) the active list row. Focus on the column labeled "S." Enter the count for "S first",61 and the count for "Q first", 45, pressing [ENTER] after each entry.
   b. Place the cursor on list L5 row 1 to make L5(1) the active list row. Focus on the column labeled "Q." Enter the count for "S first",31 and the count for "Q first", 53, pressing [ENTER] after each entry.

2. Calculate the column percents for "S".

   On the homescreen, press [2nd] [L3] [÷] [2nd] [LIST] [►] [►] selecting the [LIST] MATH menu. Select 5: sum( and press [2nd] [L3] [)] [×] 100 [STO→] [2nd] [L4], as shown in Figure 4.4.

3. Calculate the column percents for "Q."

On the homescreen, press 2nd L5 ÷ 2nd LIST ► ► selecting the LIST MATH menu. Select 5: sum( and press 2nd L5 ) × 100 STO→ 2nd L6, as shown in Figure 4.5.

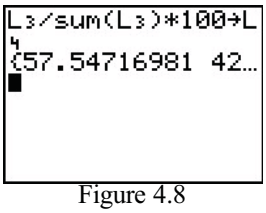


Figure 4.8

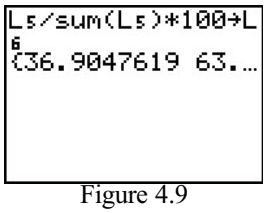


Figure 4.9

4. Select STAT ENTER to select the STAT list editor.

a. Place the cursor on list L4 row 1 to make L4(1) the active list row. View the column percents for ”S” displayed in list L4, as shown in Figure 4.10.
b. Place the cursor on list L6 row 1 to make L6(1) the active list row. View the column percents for ”Q” displayed in list L6, as shown in Figure 4.11.

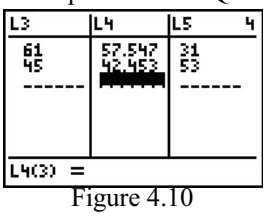


Figure 4.10

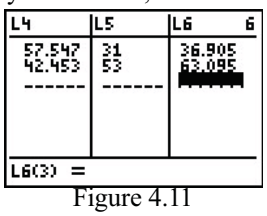


Figure 4.11

As a result of the calculation of the column percents, you are now able to add the column percents to the table of the number of occurrences of the letter ”Q” and the letter ”S”. This information is shown in Table 4.5. The % of Column for Total were calculated seperately.

| | Letters picked | | |
|---|---|---|---|
| Form | S | Q | Total |
| S first | 61 | 31 | 92 |
| % of Column | 57.5 | 36.9 | 48.4 |
| Q first | 45 | 53 | 98 |
| % of Column | 42.5 | 63.1 | 51.6 |
| Total | 106 | 84 | 190 |
| | 100.0 | 100 | 100.0 |

Table 4.5 Occurrences of the letters ”S” and ”Q”

We can use column percents to compare the rates of the form received by those who picked the letter ”S” and those who picked the letter”Q.” The first column of the table gives the data for those who picked the letter ”S.” Among the 106 individuals who picked the letter ”S”, 57.5% received the form ”S first”, and 42.5% received the form ”Q first.” The second column of the table gives the data for those who picked the letter ”Q.” Among

the 84 individuals who picked the letter "Q", 36.9% received the form "S first", and 63.1% received the form "Q first." The difference between the two sets of column percents appears to indicate a relationship. There is a relationship between two categorical variables forming a two-way table if two or more columns have different distributions of column percents.

## 4.4 Assessing the Statistical Significance of a 2×2 Table

**Example - Continued**

Question 3 in the class survey described in Section 2.1 asked 92 college students to "Randomly pick a letter-S or Q." Another 98 college students were asked to "Randomly pick a letter-Q or S." The data is contained in the *PennState1* worksheet and is displayed in Table 4.1 and Table 4.2. Table 4.3 contains the frequency table of the number of occurrences of the letter ''Q" and the letter "S''.

The steps involved in computing a chi-square-test on the TI-83 S.E. and TI-84 S.E. require:
entering the observed counts in a matrix. Enter that matrix variable name at the Observed: prompt in the $\chi^2$.Test editor with the default being matrix A: [A]. At the Expected: prompt, enter the matrix variable name to which you want the computed expected counts to be stored with the default being matrix B: [B].
Calculating the $\chi^2$ test statistic.
Examining the matrix of Expected counts obtained by calculating the $\chi^2$ test statistic.

**Follow these steps** to conduct a chi-square test, including observed counts, computing a chi-square statistic and finding the p-value.

1. Enter the observed counts in matrix A.
   a. Press [2nd] [MATRIX], located on row 4, left hand column. and press [►] [►] selecting the [MATRIX] EDIT menu, as shown in Figure 4.12. Select matrix [A]. Press [ENTER].
   b. Enter the dimensions of matrix A. Press 2 (rows), [ENTER], press 2 (columns) [ENTER]. Refer to Table 4.3 to obtain the observed counts. Enter the element in the first row, first column: 61; press [ENTER]. Enter the element in the first row, second column: 31; press [ENTER]. Enter the element in the second row, first column: 45; press [ENTER]. Enter the element in the second row, second column: 53; press [ENTER]. The resulting matrix is

shown in Figure 4.13.

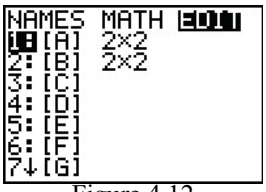


Figure 4.12

Figure 4.13

Press 2nd QUIT.

2. Compute the chi-square statistic.

Press STAT ► ► selecting the STAT TEST menu. Select C: $\chi^2$ Test, as shown in Figure 4.14. Press ENTER. If matrix A is not listed in Observed: and matrix B is not listed in Expected,as shown in Figure4.15 then follow these instructions:

a. Place matrix A in Observed: by selecting 2nd MATRIX >NAMES 1: [A] and pressing ENTER , as shown in Figure 4.16.

b. Place matrix B in Expected: by selecting 2nd MATRIX >NAMES 2: [B] and pressing ENTER, as shown in Figure 4.16.

Figure 4.14

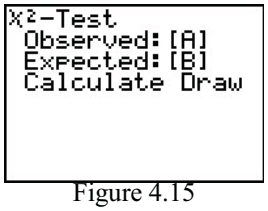


Figure 4.15

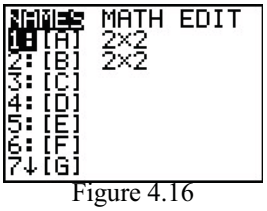


Figure 4.16

c. Calculate the chi-squared statistic by highlighting Calculate and press ENTER, as shown in Figure 4.17. The results are shown in Figure 4.18.

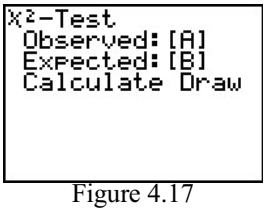


Figure 4.17

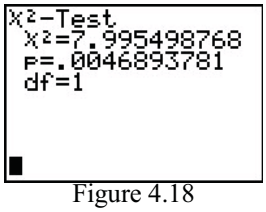


Figure 4.18

Using the TI calculator, the p-value is found to be 0.005. The p-value tells us that the chance is only 0.005 (which is really 5 in 1000) that we would get a chi-square statistic as large as 7.995 (or larger) if there really is no relationship between the order of the letters on the form and the letter that would be picked by people in this population. In the context of this problem, this means that there is a statistically significant relationship between the form of the question (”S first” or ”Q first”) and the letter picked *in the population*.

d. View the expected counts for the two-way table..

Press 2nd MATRIX >NAMES 2: [B] and pressing ENTER to view the expect counts shown in Figure 4.19.

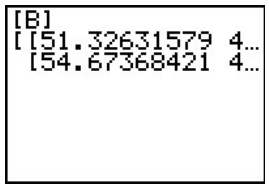



Figure 4.19

Use the right arrow key, ▶, to scroll through the expected values. As a result of viewing the expected values, you are now able to add the expected values to the table of the number of occurrences of the letter "Q" and the letter "S". This information is shown in Table 4.6. The % of Column for Total were calculated seperately.

| | Letters picked | | |
|---|---|---|---|
| Form | S | Q | Total |
| S first | 61 | 31 | 92 |
| Expected Value | 51.33 | 40.67 | |
| Q first | 45 | 53 | 98 |
| Expected Value | 54.67 | 43.33 | |
| Total | 106 | 84 | 190 |

Table 4.6 Occurrences of the letters "S" and "Q"

## The Family of $\chi^2$-Distributions

A $\chi^2$-distribution is used to find the p-value for an $\chi^2$-test of the null hypothesis that there is no association between the two variables. The family of $\chi^2$-distributions is a family of skewed distributions, each with a minimum value of 0. A specific $\chi^2$ distribution is indicated by the parameter called degreesof freedom In $\chi^2$-test,the degrees of freedom is $df = k - 1$ (number of groups - 1).

## Finding a P-value

Finding a p-value given the $\chi^2$ value and degrees of freedom. In Example 4.10, Figure 4.18, the p-value is reported as part of the output. The TI distribution function $\chi^2 cdf$ (*lowerbound*, *upperbound*, *df*) computes the $\chi^2$ distribution probability between the *lowerbound* and *upperbound* for the specified *df*.

**Follow these steps** to obtain the p-value obtained in Example 4.10, Figure 4.18.

1. Find a p-value for an $\chi^2$-distribution:

   Press 2nd DISTR, using the down arrow key, ▼, several times to select 7: $\chi^2$cdf(, as shown in Figure 4.20 and press ENTER. Type the *lowerbound*, 7.995, *upperbound*,1E99, *df*,1). The *upperbound*,1E99, is translated as $1 \times 10^{99}$. Press ENTER to execute the command. The results are shown in Figure 4.21.

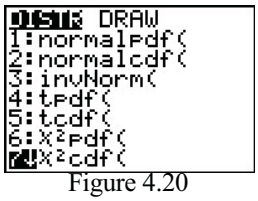


Figure 4.20

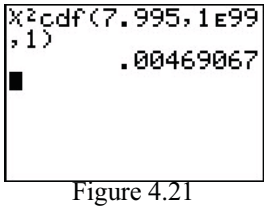


Figure 4.21

   The area to the right of $\chi^2 = 7.995$ under the $\chi^2$- distribution is the same as the p-value, 0.005.

# Chapter 5
# Sampling: Surveys and How to Ask Questions

There are two major categories of statistical techniques that can be applied to data. The first is **descriptive statistics**, in which we use numerical and graphical summaries to characterize a dataset. We partially covered descriptive statistics in Chapter 2, and we introduced additional descriptive techniques in Chapters 3 and 4. The second important category of statistical techniques is **inferential statistics**, in which we use sample data to make conclusions about a broader range of individuals than just those who are observed. For example, in Case Study 1.6 about aspirin use and the risk of heart disease, the data from a sample of 22,071 physicians was used to *infer* that taking aspirin helps prevent heart attacks for all men similar to the participants.

In Chapters 5 and 6, you will learn how to collect representative data. In these chapters you willleam that the data collection method used affects the extent to which you can use sample data to make inferences about a larger population. Descriptive summaries such as the mean and standard deviation, as well as graphical techniques, can be used whether the data are from a sample or from an entire population, but inferential methods can be used only when the data in hand are from a *representative* sample for the question being asked a about a larger population. When you use inferential methods, a key concept is that you have to think about *both* the source of the data *and* the question(s) of interest. A dataset may contain representative infonnation for some questions but not for others.

The **Fundamental Rule for Using Data for Inference** is that available data can be used to make inferences about a much larger group *if the data can be considered ,to be representative with regard to the question(s) of interest*.

After reading this chapter you should be able to:

1. Select a simple random sample.

**Keystrokes Introduced**

1. [2nd] [LIST] [►], selecting 5: seq( from the OPS menu. The arguments in seq( are seq(*expression*,*variable*,*begin*,*end* [,*increment* ]). You will use seq(*expression*,*variable*,*begin*,*end* [,*increment* ]) to create a column of ID lables.
2. [MATH] [►] [►] [►] selecting 5: randInt( from the PRB menu. The arguments in randInt( are randInt(*lower*,*upper* [,*numtrials* ]). You will use randInt( are randInt(*lower*,*upper* [,*numtrials* ]) to randomly select students.

## 5.1 Populations, Samples, and Simple Random Samples

In most statistical studies, the objective is to use a small group of units to make an inference about a larger group. The larger group of units about which inferences are to be made is called the **population**. The smaller group of units actually measured is called the **sample**. Sometimes measurements are taken on the whole group of interest, in which case *these measurements comprise a census of the whole population*. Occasionally you will see someone make the mistake of trying to use census data to make inferences to some hypothetical ”larger group” when there isn't one.

### Simple Random Samples

Rememberthe fundamental rule for making valid inferences about the group represented by the sample for which the data were measured: *The data must be representative of the larger group with respect to the question of interest*.The principal way to guarantee that sample data represents a larger population is to use a **simple random sample** from the population.

With a **simple random sample**, every conceivable group of units of the required size from the population has the same chance to be the selected sample.

An ideal data collection method is to obtain a simple random sample of the population of interest, or to collect sample data using one of the more complex random sampling methods described later in this chapter. In some research studies, however, random sampling is not possible for both practical and ethical reasons. For instance, suppose researchers want to study the effect of using marijuana to reduce pain in cancer patients. It would be neither practical nor ethical to select a random sample of all cancer patients to participate. Instead, the researchers would use volunteers who want to take part, and hope these volunteers represent the larger population of all cancer patients. The use of volunteers will be discussed more fully in Chapter 6, when we cover randomized experiments.
Simple random samples and related sampling methods *are* typically used for one type of statistical study: sample surveys or polls. Remember from Chapter 1 that in a *sample survey* the investigators gather opinions or other information from each individual included in the sample. Because this gathering of information is usually not time consuming or invasive, it is often both practical and ethical to contact a large random sample from the population of interest. Throughout this chapter we will learn more about how to select simple random samples and how to conduct sample surveys.

**Example: Finding a Simple Random Sample Using UCDavis1.**

Students in a liberal arts course in statistical literacy were given a survey that inc1uded questions on how many hours per week they watched television. The responses are shown in Table 5.1 and are contained in the UCDavis1 data file.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13 | 2 | 20 | 15 | 8 | 3 | 2 | 4 | 8 | 1 | 8 | 28 | 4 |
| 11 | 10 | 1 | 10 | 10 | 1 | 4 | 2 | 40 | 16 | 10 | 30 | 10 |
| 2 | 10 | 15 | 4 | 6 | 100 | 6 | 15 | 1 | 2 | 4 | 10 | 1 |
| 1 | 4 | 1 | 6 | 2 | 4 | 10 | 18 | 20 | 4 | 20 | 5 | 0 |
| 11 | 0 | 1 | 2 | 8 | 1 | 3 | 6 | 10 | 15 | 15 | 12 | 2 |
| 4 | 15 | 21 | 4 | 4 | 8 | 2 | 4 | 10 | 2 | 9 | 7 | 14 |
| 2 | 4 | 0 | 10 | 10 | 25 | 6 | 14 | 0 | 21 | 14 | 11 | 8 |
| 2 | 2 | 14 | 2 | 6 | 20 | 14 | 1 | 14 | 10 | 15 | 2 | 10 |
| 6 | 20 | 20 | 35 | 15 | 5 | 14 | 35 | 1 | 4 | 0 | 14 | 5 |
| 5 | 5 | 1 | 1 | 9 | 15 | 5 | 8 | 1 | 10 | 2 | 7 | 14 |
| 1 | 1 | 2 | 1 | 4 | 3 | 8 | 1 | 3 | 12 | 30 | 15 | 1 |
| 9 | 25 | 2 | 3 | 1 | 4 | 30 | 20 | 3 | 2 | 15 | 16 | 5 |
| 8 | 10 | 2 | 8 | 10 | 10 | 6 | 4 | 8 | 3 | 1 | 5 | 8 |
| 2 | 9 | 1 | 5 | | | | | | | | | |

Table 5.1

**Follow these steps** to find a simple random sample of 10 students weekly television watching amounts (variable is TV).

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the weekly television watching amounts (variable is TV) in list L3.

      Place the cursor on list L3 row 1 to make L3(1) the active list row. Enter the weekly television watching amounts from Table 5.1 row by row. type 13, 2, 20, ...9, 1, 5 pressing [ENTER] after each entry to enter all 173 weekly television watching amounts, as shown in Figure 5.1.

b. Create a column of ID lables in list L1.
Press 2nd QUIT. On the homescreen, we will create a list of ID labels from 1 to 173 in list L1. Press 2nd LIST ►, selecting 5: seq(from the OPS menu, as shown in Figure 32. Press ENTER, placing seq( on the homescreen.
The arguments in seq( are seq(*expression*,*variable*,*begin*,*end* [,*increment* ]).
Enter the expression by pressing $X,T,\theta,n$ found on row three, column two; press ,.
Enter the variable by pressing $X,T,\theta,n$, ,.
Enter the values for begin, end, and increment. Type 1,173,1 followed by ).
Store the results in list L1 by pressing STO 2nd L1. Press ENTER to execute the comand. The homescreen is shown in Figure 5.3.

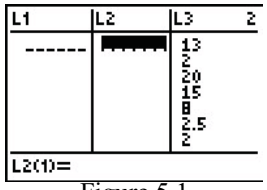


Figure 5.1

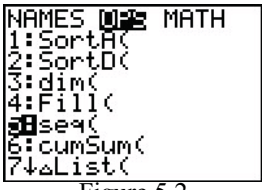


Figure 5.2

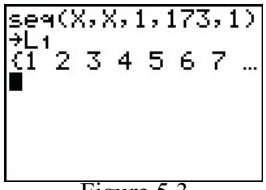


Figure 5.3

3. Randomly select 10 students weekly television watching amounts, storing the results in list L2.

On the homescreen, press MATH ► ► ► selecting 5: randInt( from the PRB menu, as shown in Figure 5.4. Press ENTER placing randInt( on the homescreen.
The arguments in randInt( are randInt(*lower*,*upper* [,*numtrials* ]).
Enter the lower, upper and numtrials by typing 1,173,10 followed by STO 2nd L2.
.Press ENTER to execute the comand. The results are shown in Figure 5.5.

4. View the randomly selected 10 students weekly television watching amounts.

Press STAT ENTER to select the STAT list editor.
The ID lables are displayed in list L1, the 10 randomly selected 1D labels are displayed in list L2, and the weekly television watching amounts are displayed in list L3, as shown in Figure 5.6.

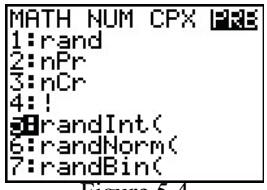


Figure 5.4

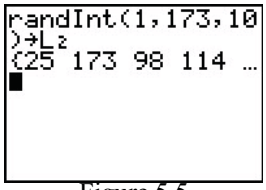


Figure 5.5

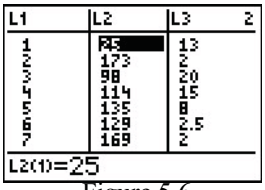


Figure 5.6

The TI output, as shown in Figure 5.6, indicates the follwing 10 randomly selected ID lables act as pointers to 10 students weekly television watching amounts: 25→ 30, 173→ 5, 98→ 14, 114→ 4, ..., 15→ 10.
Your results most certainly would be different since these are randomly selected labels.

# Chapter 6
# Gathering Useful Data For Examining Relationships

In this chapter, we learn about ways to collect data in order to examine relationships between variables. We have already seen several examples that involved possible links between variables. In Chapter 2, Example 2.1 was about the connection between gender and seat belt use for $12^{th}$ grade students. Example 2.2 was about a possible connection between the use of nightlights in infancy and nearsightedness. Case Study 1.6 described a study that demonstrated a link between taking an aspirin a day and a decreased risk of heart attacks for men.
In studies like these, we want to know if a cause-and-effect relationship exists. That is, we want to know if changing the value of one variable cause ehanges in another variable. We will learn in this chapter that the way a study is conducted affects our ability to infer that a cause-and-effect relationship exists.

## 6.1 Speaking the Language of Research Studies

Although there are a number of different strategies for collecting meaningful data, there is common terminology used in most of them. Statisticians tend to borrow words from common usage and apply a slightly different meaning, so be sure you are familiar with the special usage of a word in a statistical context.

**Types of ResearchStudies**

There are two basic types of statistical research studies conducted to detect relationships between variables:

*observational studies*
*experiments*

In an observational study, the researchers simply observe or question the participants about opinions, behaviors, or outcomes. Participants are not asked to do anything differently. For example, Case Study 1.5 described an observational study in which blood pressure and frequency of certain types of religious activity (like prayer and church attendance) were measured. The goal was to see if people with higher frequency of religious activity had lower blood pressure. Researchers simply measured blood pressure and frequency of religious activity. They did not ask participants to change how often they prayed or went to religious services, or change any other aspect of their lives.

In an experiment, researchers manipulate something and measure the effect of the

manipulation on some outcome of interest. Randomized experiments are experiments in which the participants are randomly assigned to participate in one condition or another. The different "conditions" are called treatments.

A major theme of this chapter will be that a randomized experiment provides stronger evidence of a cause-and-effect relationship than an observational study.

## 6.2 Designing a Good Experiment

An experiment measures the effect of manipulating the environment of the participants in some way. With human participants, the manipulation may include receiving a drug or medical treatment, going through a training program, agreeing to a special diet, and so on. Most experiments on humans use volunteers because you can't very well force someone to accept a manipulation. Experiments are also done on other kinds of experimental units, such as when different growing conditions are compared for their effect on plant yield, or different paints are applied on highways to see which ones last longer. The idea is to measure the effect of the feature being manipulated, the explanatory variable, on the response variable.

In a randomized experiment, participants usually are randomly assigned to either receive a specific treatment or to take part in a control group. The purpose of the random assignment is to make the groups approximately equal in all respects except for the explanatory variable, which is purposely manipulated. Differences in the response variable between the groups, if large enough to rule out natural chance variability, can then be attributed to the manipulation of the explanatory variable.

After reading this chapter you should be able to:

1. Use simulation to obtain a random sample.

**Keystrokes Introduced**

1. [MATH] [►] [►] [►] the PRB (probability) menu, select 5: randInt(*lower*,*upper* [,*numtrials*]). The command generates and displays a random integer within a range specified by *lower* and *upper* integer bounds for a specified number of trials *numtrials*.
2. [2nd] [LIST] [►] OPS>1: SortA(*listname*) sorts elements of *listname* in ascending order.
3. [DEL] used to delete an entry.

## 6.3 Simple Random Sampling and Randomization

We have already encountered several examples of the type of problem we will study in this chapter. In Chapter 2, for instance, we described a study of 479 children that found that children who slept either with a nightlight or in a fullylit room before the age of two had a higher incidence of myopia (nearsightedness) later in childhood. Example 6.10 - Finding Gifted ESP Participants will be used to illustrate how relationships between categorical variables may be presented.

**Example 6.3 - Assigning Children to Lift Weights**

In Case Study 6.2, 43 children were randomly assigned to one of three treatment groups. Children in group 1 performed weight-lifting repetitions with a heavy load, group 2 performedmore repetitions but with a moderate load, and group 3 was a control group that did not lift weights. There were 15 children assign to group 1, 16 to group 2, and 12 to group 3.
Suppose we are asked to randomly assign children to treatment goups. How could we carry out this randomization?
One way would be to use a random integer generator, like the one on the TI-83 Plus S.E. or TI-84S.E. First , we think of the children as being labeled with integers from 01 to 43. Then you will choose a simple random sample of size 15.
**Follow these steps** to choose 15 children to assign to Group 1 and the 12 children to assign to the Control Group.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Use the random integer function to select 15 children to assign to Group 1.
   a. On the homescreen, press [MATH], located on the fourth row, left column. Press [►] [►] [►] to select the PRB (probability) menu. Select 5: randInt( to generate and store a random integer.Type 1,43,15 to select 15 integers from 1 to 43. Press [)] [STO→] [2nd] [L1] to store the random integers in list L1, as shown in Figure 6.1.

b. Check for duplicates.

On the homescreen, press [2nd] [LIST], located on the third row, column 3. Press 2: SortA( and [2nd] [L1] [)] [ENTER] to place the list in ascending order, as shown in Figure 6.2. Choose [STAT] >EDIT to view the list in the [STAT] list editor. If duplicates exist, repeat all of Step 2, until no duplicates are present by pressing [2nd] [QUIT], exiting the [STAT] editor.Press [2nd] [ENTER] [2nd] [ENTER] to execute the commands once again.

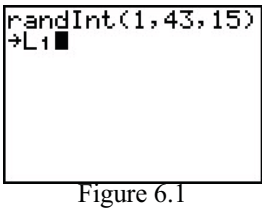


Figure 6.1

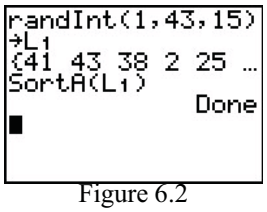


Figure 6.2

3. Use the random integer function to select 12 children to assign to the Control Group.
   a. On the homescreen, press [MATH], located on the fourth row, left column. Press [►] [►] [►] to select the PRB (probability) menu. Select 5: randInt( to generate and store a random integer.Type 1,43,30 to select 30 integers from 1 to 43. Press [)] [STO→] [2nd] [L2] to store the random integers in list L2, as shown in Figure 6.3.

   b. Check for duplicates.

   On the homescreen, press [2nd] [LIST], located on the third row, column 3. Press 2: SortA( and [2nd] [L2] [)] [ENTER] to place the list in ascending order, as shown in Figure 4.4. Choose [STAT] >EDIT to view the lists in the [STAT] list editor. View list L1 and list L2 side by side. If duplicates exist in L2, press [DEL] to delete a duplicate entry in L2. If an entry appear in list L1 and also in list L2, press [DEL] to delete the entry in list L2. Do this until only 12 entries appear in list L2. You may wish to repeat all of Step 3, until no duplicates are present in list L2 by pressing [2nd] [QUIT], exiting the [STAT] editor.Press [2nd] [ENTER] [2nd] [ENTER] to execute the commands once again. The results are shown in Figure 6.5. Your results are most likely to be different since these are random numbers.


Figure 6.3

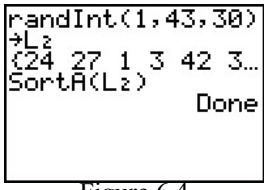



Figure 6.4

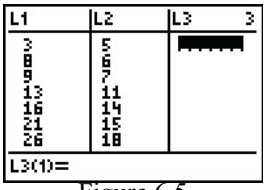



Figure 6.5

The TI calculator output, as shown in Figure 6.5, indicates 15 students labeled 3, 8, 9, ... are in Group 1, while 12 students labeled 5, 6, 7 are in the Control Group. The remaining 16 children will constitute Group 2.

Other methods of making random assignments would also work. For instance, Minitab or Excell could also be used to create random assignments.

# Chapter 7
# Probability

Statistical methods are used to evaluate information in uncertain situations and probability plays a key role in that process. Remember our definition of statistics from Chapter 1: *Statistics is a collection of procedures and principles for gathering data and analyzing information in order to help people make decisions when faced with uncertainty*. Decisions like whether to buy a lottery ticket, whether to buy an extended warranty on a computer,or which of two courses to take are examples of decisions that you may have to make that involve uncertainty and the evaluation of probabilities.

Probability calculations also are a key element of statistical inference. In Chapter 6 we introduced p-values, which are probabilities used to determine if the results of a study are statistically significant. As a reminder of how p-values are used, consider Case Study 1.6 in which 22,071 physicians were randomly assigned to take either aspirin or a placebo. There were 189 heart attacks in the placebo group but only 104 in the aspirin group. Could this have happened just by the luck of how the physicians were randomized to the treatment groups?

Suppose that regardless of which group they were in, 104 + 189 = 293 of the men would have had heart attacks anyway. What is the probability that, just by the luck of random assignment, the numbers of heart attacks in the two groups would have been so different? In other words, if aspirin and placebo are equally effective (or ineffective), what is the probability that we would see such a large discrepancy in the proportion of heart attacks in the two groups? The answer is the p-value, which is less than .00001. This is strong evidence that these results did not just occur by chance. From this, we conclude that aspirin really did reduce the number of heart attacks in the group that took it.

After reading this chapter you should be able to:

1. Use simulation to estimate probabilities.

**Keystrokes Introduced**

1. [2nd] [LIST] [►] OPS>5: seq(*expression, variable, begin, end[,increment]* ) returns a list.
2. [MATH] [►] [►] -PRB>7: randBin(*numtrials,prob,[,numsimulations]* ) generates and displays a random real number from a specified Binomial distribution.
3. [MATH] [►] [►] -PRB>6: randInt(*lower,upper,numtrials* ) generates and displays a random integer within a range specified by *lower* and *upper* for a specified number of trials.

## 7.1 Using Simulation to Estimate Probabilities

Some probabilities are so difficult or time-consuming to calculate that it is easier to simulate the situtation repeatedly using a computer or calculator and observe the relative frequency of the event of interest. If you simulate the random circumstance $n$ times and the outcome of interest occurs in $x$ out of those $n$ times, then the estimated probability for the outcome of interest is $\frac{x}{n}$. This is an estimate of the long-run relative frequency with which the outcome would occur in real life.

**Example 7.30 - Finding Gifted ESP Participants**

An ESP test is conducted by randomly selecting one of five video clips and playing it in one building, while a participant in another building tries to describe what is playing. Later, the participant is shown the five video clips and is asked to determine which one best matches the description he or she had given. By chance, the participant would get this correct with probability 1/5. Individual participants are each tested eight times, with five new video clips each time. They are identified as ”gifted” If they guess correctly at least five times out of the eight tries. Suppose people actually do have some ESPand can guess correctly with probability .30 (instead of the .20 expected by chance). What is the probability that a participant will be identified as ”gifted”?
In Chapter 8 you will learn how to solve this kind of problem, but we can simulate the answer using a TI calculator to produce the digits 0, 1,2, .. .,9 with equal likelihood. A random number table is available in the text. Many calculators and computers will simulate these digits. Here are the steps needed for one ”repetition”:
Each ”guess” is simulated with a digit, equally likely to be 0 to 9.
For each participant, we simulate eight ”guesses” resulting in a string of eight digits.
If a digit is 7, 8, or 9, we count that guess as ”correct” so P(correct) =
3/10 = .3, as required in the problem. If the digit is 0 to 6, the guess Is ”incorrect.”
(There is nothing special about 7, 8,9; we could have used any
three digits.)
If there are five or more ”correct” guesses (digits 7, 8, 9), we count that as
”gifted.”

The entire process is repeated many times, and the proportion of times the result is a ”gifted” participant is an estimate of the desired probability.

**Follow these steps** to simulate this experiment for one participant, exploring the step-by-step process.

1. Preparations:
   a. Turn off all”Y=” functions.

      Press Y= and press CLEAR to remove all functions For each line that is not blank, place the cursor on the function and press CLEAR Press

2nd QUIT.

b. Clear all lists in the Stat editor.

Press STAT, selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press ENTER to execute the command.

2. Place a sequence of random real numbers from a integer distribution into list L1.

On the homescreen press 2nd LIST ▶ to select the LIST OPS menu.

a. Obtain a sequence of 8 random digits from {0,1,2,3,4,5,6,7,8,9}.
Press MATH ▶ ▶ ▶ to the PRB menu. Select 5: randInt(. Type 0 , 9 ) x,T,$\theta$,$n$ , 1 , 8 , 1 ). The x,T,$\theta$,$n$ key is located on the 3rd row, end column. Press ENTER to execute the command, as shown in Figure 7.1.

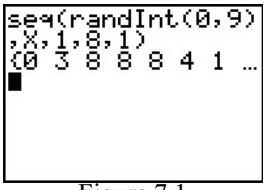


Figure 7.1

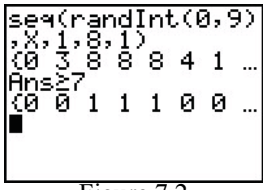


Figure 7.2

3. If a digit is 7, 8, or 9, we count that guess as "correct." Code each digit in list L1 as success (1), if the digit is a 7, 8, or 9. Code each digit in list L1 as failure (0) if the digit is less than 7. You will store the results in list L2.

Press 2nd TEST displaying the TEST menu.

a. Select 4: $\geq$ and press ENTER. Press 7 to compare to the smallest "success" number. Observe that "0" is the code for failure or False (the digit was $< 7$). Observe that "1" is the code for success or true (the digit was $\geq 7$). The output from the TI calculator is displayed in Figure 7.2.

4. Count the number of successes.

On the homescreen press 2nd LIST ▶ ▶ to select the LIST MATH menu.

(i) Select 5: sum(. Press ENTER. Press 2nd ANS ). The ANS key is located above the gray negation key on the bottom row, column four. The number of successes are displayed in Figure 7.3.

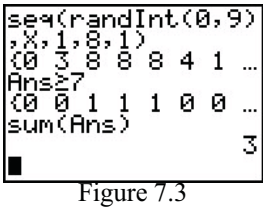


Figure 7.3

Notice that this participant got 3 "guesses" of digits that were a 7, 8, or 9.

**Follow these steps** to simulate this experiment for a 100 participants.

1. Place a sequence of random real numbers from a integer distribution into list L1.

   On the homescreen press [2nd] [LIST] [►] to select the [LIST] OPS menu.

   a. Obtain a sequence of 100 random digits {0,1,2,3,4,5,6,7,8,9}storing the sequence in list L1.
   Press [MATH] [►] [►] [►] to the PRB menu. Select 7: randBin(. Type 8 [,] 0.3 [)] [x,T,$\theta$,$n$] [,] 1 [,] 100 [,] 1 [)] [STO→] [2nd] [L1]. The [x,T,$\theta$,$n$] key is located on the 3rd row, end column. Press [ENTER] to execute the command, as shown in Figure 7.4.

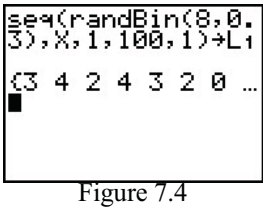


Figure 7.4

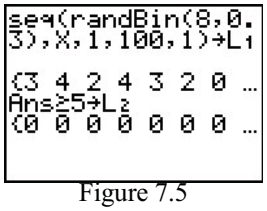


Figure 7.5

2. If there are five or more "correct" guesses (digits 7, 8, 9), we count that as "gifted." Code each "participant" in list L1 who had five or more "correct" guesses as success (1). Code each "participant" in list L1 who had fewer than five "correct" guesses as as failure (0). You will store the results in list L2.

   Press [2nd] [TEST] displaying the TEST menu.

   a. Select 4: $\geq$ and press [ENTER]. Press [5] to compare to the smallest "-success" number. Press [STO→] [2nd] [L2], storing the 1's and 0's in list L2. Observe that "0" is the code for failure or False (the digit was $< 7$). Observe that "1" is the code for success or true (the digit was $\geq 7$). The

output from the TI calculator is displayed in Figure 7.5.

3. Count the number of successes.

On the homescreen press [2nd] [LIST] [►] [►] to select the [LIST] MATH menu.

(i) Select 5: sum(. Press [ENTER]. Press [2nd] [L2] [)], selecting list L2. The number of successes are displayed in Figure 7.3.

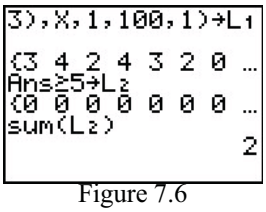



Figure 7.6

Notice that there were 2 participants who got five or more correct, so the probability of finding a ”gifted” participant in this simulation is about $\frac{2}{100} = 0.02$. In other words, if everyone is equally talented, and each guess is correct with probability 0.3, there there will be five or more correct guesses out of eight tries with probability 0.02, or about 2% of the time.

# Chapter 8
# Random Variables

The numerical outcome of a random circumstance is called a random variable. In this chapter, we'll learn how to characterize the pattern of the distribution of the values that a random variable may have, and we'll learn how to use the pattern to find probabilities. Patterns make life easier to understand and decisions easier to make. For instance, dogs come in a variety of breeds, sizes, and temperaments, but all dogs fit certain patterns that veterinarians can rely upon when treating nearly any type of dog. If a veterinarian had to learn a different pattern for treating every different breed, it might be nearly impossible for any individual to learn enough to be able to treat dogs in general.

After reading this chapter you should be able to:

1. List the probabilities for a binomial experiment.
2. Find exact and cumulative probabilities for a specific value of x, given n and p.
3. Find probabilities for a uniform distribution.
4. Find standard normal probabilities.
5. Find probabilities for any normal distribution.
6. Find percentiles for a normal distribution.

**Keystrokes Introduced**

1. 2nd DISTR, selecting 0: binomialpdf(. The arguments are binomialpdf(*numtrials*,*p*[,*x*]). The function computes a probability at $x$ for the discrete binomial distribution wiht the specified *numtrials* and probability $p$ of success on each trial.
2. 2nd DISTR, selecting A: binomialcdf( . The arguments are binomialpdf(*numtrials*,*p*[,*x*]). The function computes a cumulative probability at $x$ for the discrete binomial distribution wiht the specified *numtrials* and probability $p$ of success on each trial.
3. 2nd DISTR, selecting 2: normalcdf( . The arguments are normalcdf(*lowerbound*,*upperbound* [,$\mu$,$\sigma$]). The function computes the normal distribution probability between *lowerbound* and *upperbound* for the specified $\mu$ and $\sigma$.
4. 2nd DISTR, selecting 3: invNorm( .The arguments are invNorm(*area* [,$\mu$,$\sigma$]). The function computes the inverse cumulative normal distribution function for a given *area* under the normal distribution specified by $\mu$ and $\sigma$.

## 8.1 Binomial Random Variables

In this section, we consider an important family of *discrete* random variables called *binomial random variables*. Certain conditions must be met for a variable to fall into this family, but the basic idea is that a binomial random variable is a count of how many times an event occurs (or does not occur) in a particular number independent observations or trials that make up a random circumstance.

**Binomial Experiments and Binomial Random Variables**

The number of heads in three tosses of a fair coin, the number of girls in six independent births, and the number of men who are six feet tall or taller in a random sample of ten adult men from a large population are all examples of binomial random variables. A binomial random variable is defined as $X =$ number of successes in the n trials of a binomial experiment.

A **binomial experiment** is defined by the following conditions:

1. There are $n$ "trials"where $n$ is specified in advance and is not a random value.
2. There are two possible outcomes on each trial, called "success" and "failure" and denoted Sand F.
3. The outcomes are independent from one trial to the next.
4. The probability of a "success" remains the same from one trial to the next, and this probability is denoted by $p$. The probability of a "failure" is $1-p$ for every trial.

**Example List the Probabilities for a Binomial Experiment**

As an example of listing the probabilites for a binomial experiment, let us use $n = 10$ and $p = 0.25$ as an example.

**Follow these steps** to find the probabilities for a binomial experiment where $n = 10$ and $p = 0.25$.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.
2. Enter data using the [STAT] list editor.

Press [STAT] [ENTER] to select the [STAT] list editor.

a. Method 1: List the probability distribution in the [STAT] editor..

Enter the data for the binomial random variable in list L1.

Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter values from 0 to 10: 0, 1, 2, ...,10 pressing [ENTER] after each entry. Place the cursor at the top of list L2, on the label L2. Press [2nd] [DISTR], selecting 0: binomialpdf( , and Press [ENTER]. Type the number of trials, 10, a [,] , the probability of success, 0.25, and [)] as shown in Figure 8.2. Press [ENTER] to execute the command. The results are shown in Figure 8.3

Figure 8.1

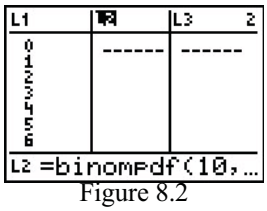


Figure 8.2

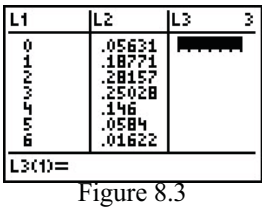


Figure 8.3

The TI output, as shown in Figure 8.3, indicates the $P(x = 0) = 0.05631$, $P(x = 1) = 0.18771$, $P(x = 2) = 0.28157$, etc.

3. Method 2: Listing the probability distribution on the homescreen..

Press [2nd] [QUIT] returning to the homescreen.
Press [2nd] [DISTR], selecting 0: binomialpdf( , type the number of trials, 10, a [,] , the probability of success, 0.25, and [)] as shown in Figure 8.4. Press [ENTER] to execute the command. The results are shown in Figure 8.5. Use the right arrow key, [▶], to view the individual probabilities.

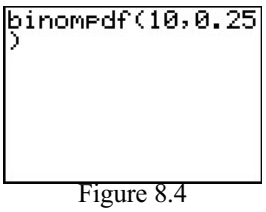


Figure 8.4

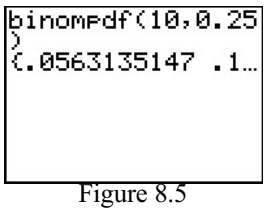


Figure 8.5

The TI output, as shown in Figure 8.5, indicates the $P(x = 0) = 0.05631$, $P(x = 1) = 0.18771$, $P(x = 2) = 0.28157$, etc.

4. Method 3: Producing individual probabilities..

Press 2nd DISTR, selecting 0: binomialpdf( , type the number of trials, 10, a , , the probability of success, 0.25, the number of successes, 0, and ) as shown in Figure 8.6. Press ENTER to execute the command. The results are shown in Figure 8.7.

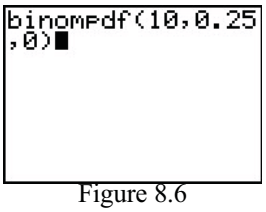


Figure 8.6

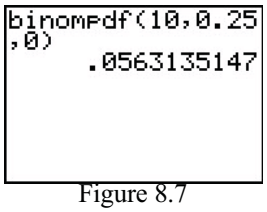


Figure 8.7

The TI output, as shown in Figure 8.7, indicates the $P(x = 0) = 0.05631$.

Repeat the above process for another value of x.
Press 2nd DISTR, selecting 0: binomialpdf( , type the number of trials, 10, a , , the probability of success, 0.25, the number of successes, 1, and ) as shown in Figure 8.8. Press ENTER to execute the command. The results are shown in Figure 8.9.

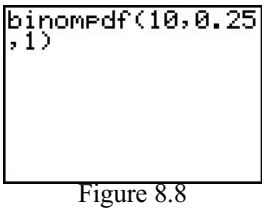


Figure 8.8

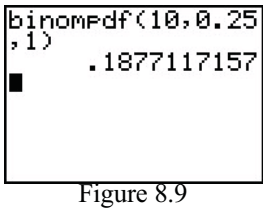


Figure 8.9

The TI output, as shown in Figure 8.9, indicates the $P(x = 1) = 0.18771$.

**Example 8.16 Calculations for Number of Girls in Ten Births**

Let X = number of girls in ten births, and assume that $p = 0.488$ is the probability that any birth is a girl. This value of $p$ is based on birth records in the United States.

**Follow these steps** to find the probability of exactly 7 girls in ten births.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press Y= and press CLEAR to remove all functions For each line that is not blank, place the cursor on the function and press CLEAR Press 2nd QUIT.
   b. Clear all lists in the Stat editor.

Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

c. Press [2nd] [DISTR], selecting 0: binomialpdf( , type the number of trials, 10, a [,], the probability of success, 0.488, the number of successes, 7, and [)] as shown in Figure 8.10. Press [ENTER] to execute the command. The results are shown in Figure 8.11.

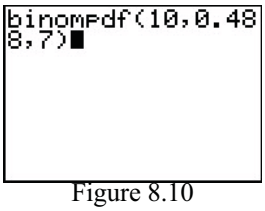


Figure 8.10

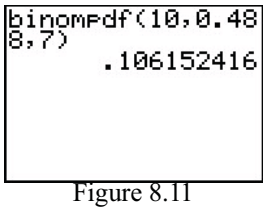


Figure 8.11

The TI output, as shown in Figure 8.11, indicates the $P(x = 7) = 0.106$.

2. Find cumulative probabilities.

a. Find the probability of having **at most 7** girls out of 10 births. An equilavent statement is to find the probability of having 7 or fewer girls out of 10 births.

Press [2nd] [DISTR], selecting A: binomialcdf( , type the number of trials, 10, a [,], the probability of success, 0.488, the number of successes, 7, and [)] as shown in Figure 8.12. Press [ENTER] to execute the command. The results are shown in Figure 8.13.

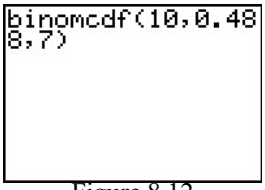


Figure 8.12

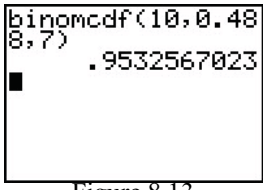


Figure 8.13

The TI output, as shown in Figure 8.13, indicates the $P(x \leq 7) = 0.9533$.

b. Find the probability of having **at least 7** girls out of 10 births. An equilavent statement is to find the probability of having 7 or more girls out of 10 births. What we will do is to subtract the probability of having at most 6 girls out of 10 births from the sum of the probabilities, 1.

Press 1 [-] [2nd] [DISTR], selecting A: binomialcdf( , type the number of trials, 10, a [,], the probability of success, 0.488, the number of successes, 6, and [)] as shown in Figure 8.14. Press [ENTER] to execute the com-

mand. The results are shown in Figure 8.15.

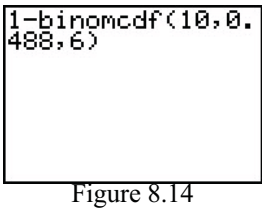



Figure 8.14

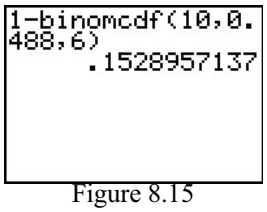



Figure 8.15

The TI output, as shown in Figure 8.13, indicates the $P(x \geq 7) = 0.1529$.

## 8.2 Continuous Random Variables

We learned in Section 8.1 that a .continuous random variable is one for which the outcome can be any value in .an interval of collection of intervals. In practice, all measurements are rounded to a specified number of decimal places, so we may not be able to accurately observe all possible outcomes of a continuous variable. Forexample, the limitations of weighing scales keep us from observing that a weight may actually be 128.3671345993pounds. Generally, however, we call a random variable a continuous random variable ifthere are a large number of observable outcomes covering an interval or set of intervals.
For a discrete random variable,we can find the probability that the variable $X$ exactly equals a specified value. We can't do this for a continuous random variable. For a continuous random variable, we are only able to find the probability that $X$ falls between two values. In other words, unlike discrete random variables, continuous random variables do not have probability distribution functions specifying the exact probabilities of specified values. Instead, they have probability density functions, which are used to find probabilities that the random variable falls into a specified interval of values.

**Example 8.19 Time Spent Waiting for a Bus** I

A bus arrives at a bus stop every 10 minutes. If a person arrives at the bus stop at a random time, how long will he or she have to wait for the next bus? Define the random variable X = waiting time until the next bus arrives. The value of X could be any value between 0 and 10 minutes, and X is a continuous random variable. (In practice, the limitations of watches would force us to round off the exact time.) Figure 8.16 shows the probability density function for the waiting time. Possible waiting times are along the horizontal axis, and the vertical axis is a density scale. The height of the "curve" is .1 for all X between 0 and 1, so the total area between 0 and 10 minutes is (10)(.1) =1.

The density function shown in Figure 8.16 is a flat line that covers the interval of possible values for X. There is a "uniformity" to this density curve in that every

interval with the same width has the same probability. A random variable with this property is called a uniform random variable and is the simplest example of a continuous random variable.

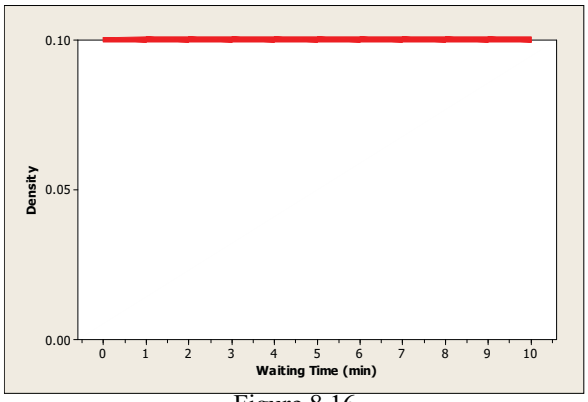



Figure 8.16

Suppose we want to find the probability that the waiting timeX was in the interval from 5 to 7 minutes. The general principle for any continuous random variable is that the probability $P(a \leq x \leq b)$ is the "area under the curve" over the interval from a to b. In this example, the "area under the curve" is the area of a rectangle that has width = 7 - 5 = 2 minutes and height = .1. This area is (2)(.1) = .2, which is the probability that the waiting time is between 5 and 7 minutes. In Figure 8.17, the shaded area represents the desired probability.

**Follow these steps** to find probabilities for a uniform distribution.

1. On the homescreen, multiply the base times the height by typing 2 [×] 0.1. Press [ENTER] to obtain the area (probability) under the uniform distribution of 0.2, as shown in Figure 8.18.

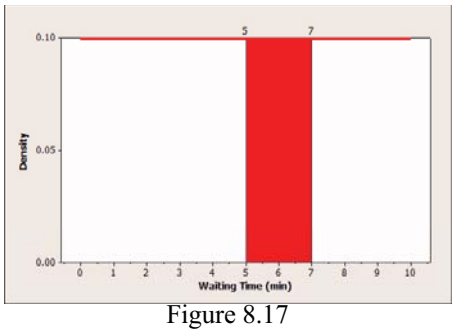



Figure 8.17

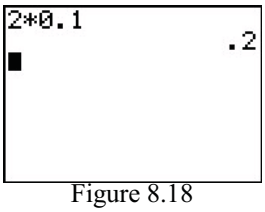



Figure 8.18

## 8.3 Normal Random Variables

The most commonly encountered type of continuous random variable is the **normal random variable** which has a specific form of a bell-shaped probability density curve called a **normal curve**. A normal random variable is also said to have a **normal distribution**. Any'normal random variable is completely characterized by specifying values for its mean, $\mu$, and standard deviation $\sigma$.
Nature provides numerous examples of measurements that follow a normal curve. The fact that so many different kinds of measurements follow a normal curve is not surprising. On many attributes, the majority of people are somewhat close to average, and as you move further from the average, either above or below, there are fewer people with such values.

**Features of Normal Curves and Normal Random Variables**

As with any continuous random variable, the probability that a normal random variable falls into a specified interval is equivalent to an area under its density curve. Also, $P(X = k) = 0$, meaning that the probability is 0 that a normal random variable $X$ exactly equals any specified value.
Some features shared by all normal curves and normal random variables ($X$) are:

1. The normal curve is symmetric and bell-shaped (but not all symmetric bell-shaped density curves are normal curves).
2. $P(X \leq \mu) = P(X \geq \mu) = 0.5$, meaning that there are equal probabilities for a measurement being less than the mean and greater than the mean.
3. $P(X \leq \mu - d) - P(X \geq \mu - d)$ for any positive number $d$. This means that the probability that $X$ is more than $d$ units below the mean equals the probability that X is more than $d$ units above the mean.
4. The Empirical Rule holds:
   a. $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$
   b. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$
   c. $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$

**Standardized Scores**

We' learned in Chapter 2 that a standardized score, also called a z-score, is the distance between a specified value and the mean, measured in number of standard deviations. We repeat the definition here using notation for random variables.

The formula for converting any value x to a z-score is

$$z = \frac{\text{Value - Mean}}{\text{Standard deviation}} = \frac{x - \mu}{\sigma}$$

## Finding Probabilities for z-Scores

A normal random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$ is said

to be a standard normal random variable and to have a standard normal distribution. When we convert values for any normal random variable to $z$-scores, it is equivalent to converting the random variable of interest to a standard normal random variable. We use the letter $Z$ to represent a standard normal random variable.

**Find a probability under the normal curve.**

**Follow these steps** to find the probability under the normal curve that z is:

1. greater than 1.31.
2. les than 1.31

3. On the homescreen, press 2nd DISTR, selecting 2: normalcdf( , type the *lowerbound* of 1.31, a , , and the *upperbound* 1 2nd EE 99, as shown in Figures 8.19 and 8.20. The key EE is located above the , on the sixth row, second column. Press ENTER to execute the command. The results are shown in Figure 8.20.

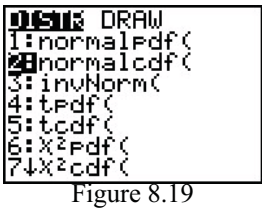


Figure 8.19

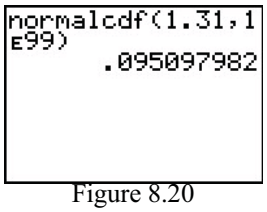


Figure 8.20

The $P(z > 1.31) = 0.0951$.

4. Find the probability under the normal curve that z is less than 1.31.

   On the homescreen, press 2nd DISTR, selecting 2: normalcdf( , type the *lowerbound* of $-1$ 2nd EE 99, a , , and the *upperbound* 1.31 as shown in Figures 8.21 and 8.22. Be sure to use the grey negation key found on the bottom row, column four. The key EE is located above the , on the sixth row, second column. Press ENTER to execute the command. The results are shown in Figure 8.22.

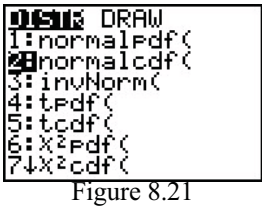


Figure 8.21

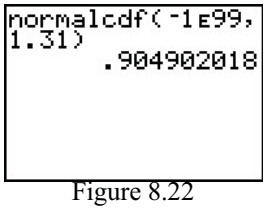


Figure 8.22

The $P(z < 1.31) = 0.9049$.

## How to Solve General Normal Curve Problems

The TI-83 SE and TI-84 SE calculators can be used to find probabilities for any general normal random variable. An important fact about nonnal random variables is that any probability problem about a normal random variable can be converted to a problem about a standard normal variable.

**Example 8.24 - Probability That Height Is Less Than 62 Inches** Assume that the heights of college women follow a normal curve with $\mu = 65$ inches and $\sigma = 2.7$ inches, we can find probabilities associated with any possible range of heights. For example, what is the probability that a randomly selected college woman is 62 inches or shorter? Equivalently, what proportion of college women are 62 inches or shorter?

**Follow these steps** to find the probability that a randomly selected college woman is 62 inches or shorter.

1. Method 1: Transform the observation to a $z - score$.
$$P(x \leq 62) = P\left(Z \leq \tfrac{62-65}{2.7}\right) = P\left(Z \leq -1.11\right)$$

   Find the probability under the normal curve that z is less than $-1.11$.

   On the homescreen, press 2nd DISTR, selecting 2: normalcdf( , type the *lowerbound* of $-1$ 2nd EE 99, a , , the *upperbound* $-1.11$, and ) as shown in Figures 8.21 and 8.22. Be sure to use the grey negation key found on the bottom row, column four. The key EE is located above the , on the sixth row, second column. Press ENTER to execute the command. The results are shown in Figure 8.22.

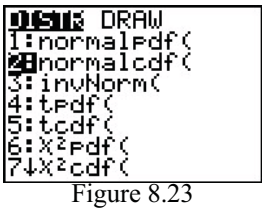


Figure 8.23

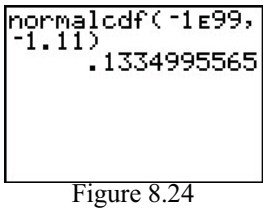


Figure 8.24

   The $P(x \leq 62) = P\left(Z \leq \tfrac{62-65}{2.7}\right) = P\left(Z \leq -1.11\right) = 0.1335$. In other words, about 13% of college women are 62 inches or shorter.

2. Method 2: Enter the lowerbound, upperbound, $\mu$, and $\sigma$ in terms of the x variable.

   On the homescreen, press 2nd DISTR, selecting 2: normalcdf( , type the *lowerbound* of $-1$ 2nd EE 99, a , , the *upperbound* 62, the value for $\mu$, 65, the value for $\sigma$, 2.7, and ) as shown in Figures 8.25 and 8.26. Be sure

to use the grey negation key found on the bottom row, column four. The key [EE] is located above the [,] on the sixth row, second column. Press [ENTER] to execute the command. The results are shown in Figure 8.26.

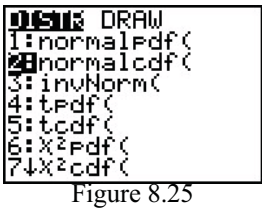


Figure 8.25

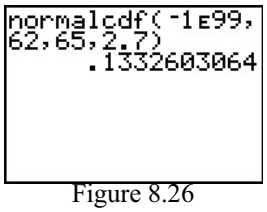


Figure 8.26

The $P(x \leq 62) = 0.1333$. In other words, about 13% of college women are 62 inches or shorter. Observe that Method 2 is good only for the $z = \frac{x-\mu}{\sigma}$ formula and is not valid for any other $z$ formula.

**Example 8.2 - Proportion of Women Who Are Taller Than 68 Inches** If we assume that college women's heights follow a normal curve with $\mu = 65$ inches and $\sigma = 2.7$ inches, we can find probabilities associated with any possible range of heights. Suppose we want to find the proportion of college women who are taller than 68 inches.

**Follow these steps** to find the proportion of college women who are taller than 68 inches.

1. Method 1: Transform the observation to a $z - score$.

$$P(x > 68) = P\left(Z > \tfrac{68-65}{2.7}\right) = P\left(Z > 1.11\right)$$

Find the probability under the normal curve that z is more than 1.11.

On the homescreen, press [2nd] [DISTR], selecting 2: normalcdf( , type the *lowerbound* of 1.11 a [,], the *upperbound* of 1 [2nd] [EE] 99, and [)] as shown in Figures 8.27 and 8.28. The key [EE] is located above the [,] on the sixth row, second column. Press [ENTER] to execute the command. The results are shown in Figure 8.28.

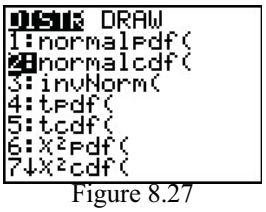


Figure 8.27

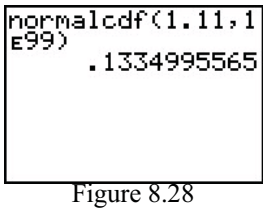


Figure 8.28

The $P(x > 68) = P\left(Z > \tfrac{68-65}{2.7}\right) = P\left(Z > 1.11\right) = 0.1335$. In other words, about 13% of college women are 68 inches or taller.

2. Method 2: Enter the lowerbound, upperbound, $\mu$, and $\sigma$ in terms of the x variable.

   On the homescreen, press 2nd DISTR, selecting 2: normalcdf( , type the *lowerbound* of 68, a , , the *upperbound* 1 2nd EE 99, the value for $\mu$, 65, the value for $\sigma$, 2.7, and ) as shown in Figures 8.29 and 8.30. The key EE is located above the , on the sixth row, second column. Press ENTER to execute the command. The results are shown in Figure 8.30.

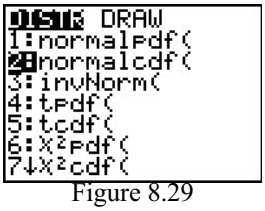


Figure 8.29

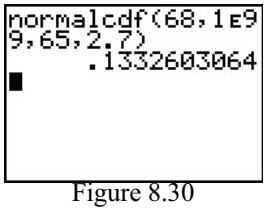


Figure 8.30

   The $P(x > 68) = 0.1333$. In other words, about 13% of college women are 68 inches or taller. Observe that Method 2 is good only for the $z = \frac{x-\mu}{\sigma}$ formula and is not valid for any other $z$ formula.

**Example 8.24 - Continued Proportion of Women Between 62 and 68 Inches Tall** If we assume that college women's heights follow a normal curve with $\mu = 65$ inches and $\sigma = 2.7$ inches, we can find probabilities associated with any possible range of heights. Suppose we want to find the proportion of college women who between 62 and 68 inches tall.

**Follow these steps** to find the proportion of college women who are taller than 68 inches.

1. Method 1: Transform the observations to $z - scores$.

$$P(x \leq 62) = P\left(Z \leq \tfrac{62-65}{2.7}\right) = P\left(Z \leq -1.11\right)$$
$$P(x \geq 68) = P\left(Z \geq \tfrac{68-65}{2.7}\right) = P\left(Z \geq 1.11\right)$$

   Find the probability under the normal curve that z is between $z = -1.11$ and $z = 1.11$.

   On the homescreen, press 2nd DISTR, selecting 2: normalcdf( , type the *lowerbound* of $-1.11$ a , , the *upperbound* of 1.11, and ) as shown in Figures 8.31 and 8.32. The key EE is located above the , on the sixth row, second column. Press ENTER to execute the command. The results are shown

in Figure 8.32.

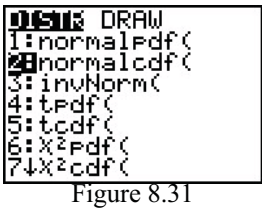



Figure 8.31

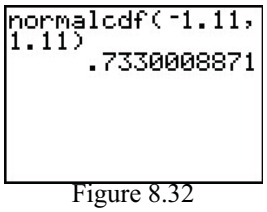



Figure 8.32

The $P(62 \leq x \leq 68) = P\left(\frac{62-65}{2.7} \leq Z \leq \frac{68-65}{2.7}\right) = P\left(-1.11 \leq Z \leq 1.11\right) = 0.7330$. In other words, about 73% of college women are between 62 and 68 inches tall.

2. Method 2: Enter the lowerbound, upperbound, $\mu$, and $\sigma$ in terms of the x variable.

   On the homescreen, press 2nd DISTR, selecting 2: normalcdf( , type the *lowerbound* of 62, a , , the *upperbound* 68, the value for $\mu$, 65, the value for $\sigma$, 2.7, and ) as shown in Figures 8.33 and 8.34. The key EE is located above the , on the sixth row, second column. Press ENTER to execute the command. The results are shown in Figure 8.34.

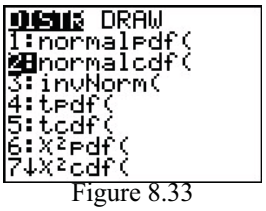



Figure 8.33

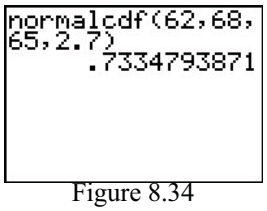



Figure 8.34

The $P(62 \leq x \leq 68) = 0.7335$. In other words, about 73% of college women are between 62 and 68 inches tall. Observe that Method 2 is good only for the $z = \frac{x-\mu}{\sigma}$ formula and is not valid for any other $z$ formula.

## 8.4 Finding Percentiles

In some problems, we want to know what value of a variable has a given percentile ranking. For example, we may want to know what pulse rate is the 25th percentile of pulse rates for men. Notice that the word percentile refers to the value of a variable. The percentile rank corresponds to the cumulative probability (area to the left under the density curve) for that value.
Suppose that the 25th percentile of pulse rates for adult males is 64 beats per minute. This means that 25% of men have a pulse rate below 64. The percentile is 64 beats per minute (a value of the variable) and the percentile rank is 25% or .25 (a cumulative probability).

**Example 8.26 - The 75th Percentile of Systolic Blood Pressures** Suppose that the blood pressures of men aged 18 to 29 years old can be described with a normal curve having mean $\mu = 120$ and standard deviation $\sigma = 10$. What is the 75th

percentile? In other words, what is the blood pressure value $x$ such that $P(Blood\ pressure \leq x) = 0.75$?

**Follow these steps** to find the 75th percentile of systolic blood pressures..

1. Method 1: Find the value of $z^*$ for which $P(Z \leq z^*) = p$.

   In order to find the value of $z^*$ for which $P(Z \leq z^*) = p$, we use the invNorm function requiring the area to the left of $z^*$. Therefore for the 75th percentile, the area to the left of the 75th percentile is $0.75$, as shown in Figure 8.35.. Take the following steps:
   On the homescreen, press 2nd DISTR, located on the fourth row, column four, above VARS. Select 3: invNorm( .Type 0.75 ) ENTER, as shown in Figure 8.36. The results are shown in Figure 8.37, indicating the appropriate $z^*$ for the 75th percentile is $0.67$, rounded to 2 decimal places.

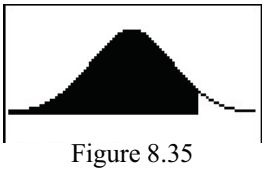


Figure 8.35

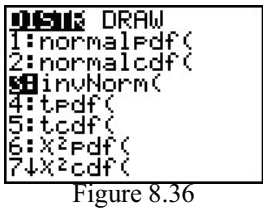



Figure 8.36

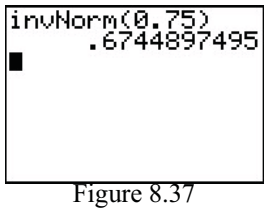



Figure 8.37

   On the homescreen, compute $x = z^*\sigma + \mu$. Type 0.67 × 10 + 120, pressing ENTER to execute the command. The results are shown in Figure 8.38.

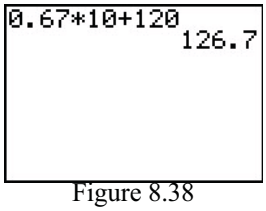



Figure 8.38

   The 75th percentile is 126.7 or about 127. $P(Blood\ pressure \leq 126.7) = 0.75$. In other words, about 75% of the blood pressures of men aged 18 to 29 years old are below 127.

2. Method 2: Enter the *area*, $\mu$, and $\sigma$ in terms of the x variable.

   On the homescreen, press 2nd DISTR, selecting 3: invNorm( .Type the area, 0.75, the value for $\mu$, 120, the value for $\sigma$, 10, and ) as shown in Figures 8.38 and 8.39. Press ENTER to execute the command. The results are shown in

Figure 8.39.

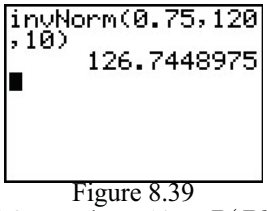



Figure 8.39

The 75th percentile is 126.7 or about 127. $P(Blood\ pressure \leq 126.7) = 0.75$. In other words, about 75% of the blood pressures of men aged 18 to 29 years old are below 127.

# Chapter 9
# Understanding Sampling Distributions: Statistics as Random Variables

This chapter introduces the reasoning that allows researchers to make conclusions about entire populations using relatively small samples of individuals. The secret to understanding how things work is to understand what kind of dissimilarity we should expect to see among different samples from the same population.
This chapter serves as an introduction to the reasoning that allows researchers to make conclusions about entire populations on the basis of a relatively small sample of individuals. The basic idea is that we must work backwards, from a sample to a population. We start with a question about a population like: How many teenagers are infected with HIV? At what average age do left-handed people die? What is the average income of all students at a large university? We collect a sample from the population about which we have the question and measure the variable of interest. We can then answer the question of interest for the sample. Finally, based on statistical theory, we will be able to determine how close our sample answer is to what we really want to know, the true answer for the population.

After reading this chapter you should be able to:

1. Simulate the sampling distribution for a sample proportion.
2. Simulate the sampling distribution for a sample mean.
3. Determine areas and probabilities for a Student's t-distribution.

**Keystrokes Introduced**

1. [2nd] [LIST] [►] OPS>5: seq(*expression, variable, begin, end[,increment]* ) returns a list.
2. [MATH] [►] [►] -PRB>7: randBin(*numtrials,prob,[,numsimulations]* ) generates and displays a random real number from a specified Binomial distribution.
3. [MATH] [►] [►] -PRB>6: randNorm($\mu$,$\sigma$,*numtrials*) generates and displays a random real number from a Normal distribution specified by $\mu$ and $\sigma$ for a specified number of trials.
4. [2nd] [DISTR] 5: tcdf(*lowerbound,uppebound,df*) computes the Student's t-distribution probability between *lowerbound* and *upperbound* for.the specified *df* (degrees of freedom).

## 9.1 Sampling Distribution for One Sample Proportion

In this Section we cover sampling distributions for one sample proportion. However, the module includes substantial discussion and explanation that should help you understand sampling distributions in general.
Suppose we conduct a binomial experiment with n trials and get successes on $x$ of the trials. Or, suppose we measure a categorical variable for a representative sample of 11individuals, and $x$ of them have responses in a certain category. In each case, we can compute the statistic $\widehat{p}$= the sample proportion $= \frac{x}{n}$, the proportion of trials resulting in success, or the proportion in the sample with responses in the specified category. If we repeated the binomial experiment or collected a new sample, we would probably get a different value for the sample proportion.
A result given in Section 8.7 of the text is that with sufficiently large n, a binomial random variable is also approximately a normal random variable. A binomial random variableX counts the number of times an event happens in n trials, but the approximate normality also applies to the proportion, $\widehat{p} = \frac{x}{n}$. Dividing each possible value of X by the sample size $n$ does not change the shape of the $n$ distribution of possible values. In other words, the sampling distributionfor a sample proportion is approximately a normal distribution.

**Example 9.4 - Possible Sample Proportions Favoring a Candidate**

This sample size in this example has been changed from 2400 in the text to a sample size of 24 in order to make this practical for a TI calculator.

Suppose that of all voters in the United States, 40% are in favor of Candidate X for president. Pollsters take a sample of 24 voters. What proportion of the sample would be expected to favor Candidate X? The rule tells that that the proportion of the sample who favor Candidate X is a random variable that has a normal distribution. The mean and standard deviation for the distribution are:

$$\text{Mean} = p = 0.40 \text{ (40\%)}$$
$$\text{s.d.}(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{24}} = 0.1$$

**Follow these steps** to simulate the sampling distribution for this sample proportion.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4,

L5, L6. Press [ENTER] to execute the command.

2. Place a sequence of random real numbers from a binomial distribution into list L1.

   On the homescreen press [2nd] [LIST] [►] to select the [LIST] OPS menu.

   a. Select 5: seq( .
   Press [MATH] [►] [►] [►] to the PRB menu. Select 7: randBin(. Type 24 [,] 0.4 [)] [x,T,$\theta$,$n$] [,] 1 [,] 100 [,] 1 [)] [STO→] [2nd] [L1]. The [x,T,$\theta$,$n$] key is located on the 3rd row, end column. Press [ENTER] to execute the command, as shown in Figure 9.1.

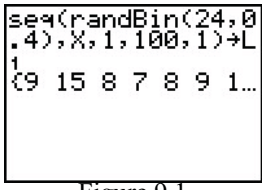


Figure 9.1

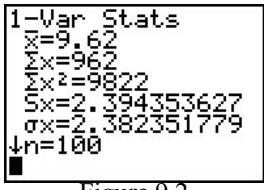


Figure 9.2

3. Obtain the numerical summaries of the number of voters in favor of Candidate X from the sample of 24 voters.

   Press [STAT] [►] to obtain the [STAT] CALC menu.

   a. Select 1: 1-Var Stats and press [ENTER]. Press [2nd] [L1] to select the number of voters in favor of Candidate X. The output from the TI calculator is displayed in Figure 9.2.
   Observe that in this random sample the mean of the number of voters in favor of Candidate X is 9.62 and the standard deviation is 2.39. The random sample that you produce may have a different mean and standard deviation.

4. Set up the plot for the histogram of the right handspan measurements for the females.

   Press [2nd] [STAT PLOT] accessing the StatPlot menu.

   (i) Press [ENTER], selecting Plot 1. Place the cursor on ON and press [ENTER]. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram. Press [ENTER]. Use the down arrow key to select L1 as the list, [2nd] [L1]. Use the down arrow key to enter 1 as the Freq:. The settings for Plot 1 are shown in Figure 9.3.

5. Enter the function to superimpose the normal curve on the histogram.

   Press [Y=], row 1, column 1, to enter the function, as shown in Figure 9.4. Press $(30/2.39\sqrt{(2\pi)})e\hat{}((-1/2)(x-9.62)^2/2.39^2)$. Observe that in this random sample the mean of the number of voters in favor of Candidate X is 9.62 with a standard deviation is 2.39, are entered into the function to determine the y-values of the graph. The 30 is a scaling factor designed to make the plot of the histogram and the normal curve coincide. Other scaling factors can be explored. You may choose to replace the mean of 9.62 and the standard deviation of 2.39 with the mean and standard deviation from your random sample. The left and right parentheses are located on row 6. Press [2nd] $\pi$, $\pi$ is located on the 5th row, right column above the [^] key. Press [2nd] $e$, $e$ is located on the 8th row, left column above the [LN] key. Be sure to use the grey negation key when you enter $(-1/2)$. The function is shown in Figure 9.4.

6. Set the Window viewing variables in order to view the graph.

   Press [WINDOW], row 1, column 2. Set Xmin to 11, Xmax to 27; Xscl to 1; Ymin to -5, being sure to use the grey negation key. Set Ymax to 31; Yscl to 1; Xres to 1. These settings are illustrated in Figure 9.5

7. View the graph.

   Press [GRAPH], to view the graph, as shown in Figure 9.6.

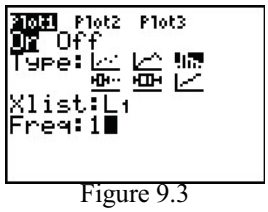


Figure 9.3

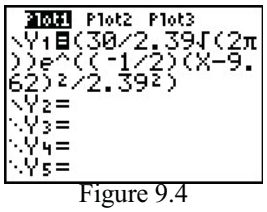


Figure 9.4

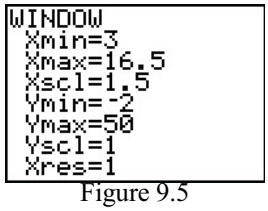


Figure 9.5

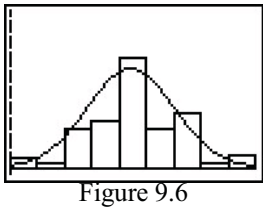


Figure 9.6

8. Turn off all plots and return the graph window to standard viewing.

   Press [2nd] [STAT PLOT], selecting PlotsOff and press [ENTER]. Press [ZOOM] and select 6: ZStandard to restore the default graph window settings.

9. Clear the function.

   Press [Y=], and Press [Y=] and press [CLEAR] to remove all functions For

each line that is not blank, place the cursor on the function and press CLEAR
Press 2nd QUIT.
Figure 9.6 indirectly shows how a sampling distribution provides information about the accuracy of a sample statistic. In that example, we learned that with a sample size of n =24 voters it is nearly certain that the proportion of voters favoring Candidate X in the sample will be within $\pm 3(0.1) = \pm 0.3$ of the true population proportion.

## 9.2 Sampling Distribution for One Sample Mean

In this Section we cover sampling distributions for one sample mean. However, the module for one mean includes substantial discussion and explanation that should help you understand sampling distributions in general.
Suppose a population consists of thousands or millions of individuals, and we are interested in estimating the mean of a quantitative variable. It we sample 25 peopleandcompute the mean of the variable for that sample,how close will that sample mean be to the population mean we are trying to estimate? Each time we take a sample we will get a different sample mean. Can we say anything about what we expect those means to be?
For example, suppose we are interested in estimating the average weight loss for everyone who attends a national weight-loss clinic for ten weeks. Suppose, unknown to us, the distribution of weight losses for everyone in this population is approximately normal with a mean of 8 pounds and a standard deviation of 5 pounds.

**Conditions for the Sampling Distribution of the Mean to Be Approximately Normal**
As with sample proportions, statisticians understand what to expect for the possible distribution of sample means in repeated sampling from the same population. Technically called the sampling distribution of the sample mean, we call this rule the Normal Curve Approximation Rule for Sample Means, or simply the Rule for Sample Means to convey what it says. Unlike the equivalent rule for proportions, it is not always necessary to have a large sample for this rule to work.If the populationof measurementsis bell-shaped,then the result holds for all sample sizes.The Rule for Sample Means applies in both of the following types of situations:

*Situation 1* The population of the measurements of interest is bell-shaped and-a-random sample of any size is measured.

*Situation 2* The population of measurements of interest is not bell-shaped, but a large random sample is measured.

*definition* The Normal Curve Approximation Rule for Sample Means can be defined as follows:

Let $\mu$ = mean for the population of interest.
Let $\sigma$ =standard deviation for the population of interest.
Let $\overline{x}$ = mean for the sample = sample mean.

If numerous random samples of the same size $n$ are taken, the distribution of possible values of X is approximately normal, with

Mean=$\mu$

Standarddeviation= s.d.($\overline{x}$) = $\dfrac{\sigma}{\sqrt{n}}$

This approximate normal distribution is called the sampling distribution of $\overline{x}$ or the sampling distribution of the mean.
*Technical Note:* The n observations in each sample must all be independent, which they will be if random samples are used.

**Example 9.7 - Hypothetical Mean Weight Loss**

For our hypothetical weight-loss example, the population mean and standard deviation were $\mu$= 8 pounds and $\sigma$ = 5 pounds, respectively, and we were taking random samples of size 25. The mean and standard deviation for the distribution are:

Mean = $\mu = 8$ pounds
s.d.$(\widehat{x}) = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1.0$

**Follow these steps** to simulate the sampling distribution for this mean.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Place a sequence of random real numbers from a normal distribution with mean = $\mu = 8$ pounds and s.d.$(\widehat{x}) = 1.0$ into list L1.

   On the homescreen press [MATH] [►] [►] [►] to the PRB menu.

a. Select 6: randNorm( . Enter the mean = $\mu = 8$, s.d.$(\widehat{x}) = 1.0$, and the number of samples, $n = 500$: 8 , 1 , 500 ) STO→ 2nd L1. Press ENTER to execute the command, as shown in Figure 9.7.

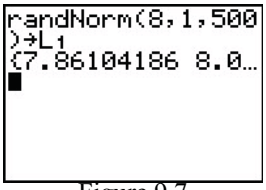


Figure 9.7

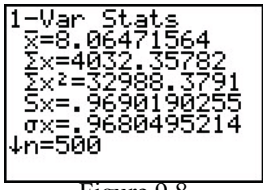


Figure 9.8

3. Obtain the numerical summaries of the sampling distribution of these sample means.

   Press STAT ▶ to obtain the STAT CALC menu.

   a. Select 1: 1-Var Stats and press ENTER. Press 2nd L1 to select the means stored in list L1. The output from the TI calculator is displayed in Figure 9.8.
   Observe that in this random sample the mean of the means is 8.06 and the standard deviation is 0.969, reasonably close to the theoretical values. The random sample that you produce may have a different mean and standard deviation.

4. Set up the plot for the histogram of the sampling distribution of the sample means.

   Press 2nd STAT PLOT accessing the StatPlot menu.

   (i) Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram. Press ENTER. Use the down arrow key to select L1 as the list, 2nd L1. Use the down arrow key to enter 1 as the Freq:. The settings for Plot 1 are shown in Figure 9.9.

5. Enter the function to superimpose the normal curve on the histogram.

   Press Y=, row 1, column 1, to enter the function, as shown in Figure 9.10. Press $(80/0.97\sqrt{(2\pi)})e\hat{}((-1/2)(x-8.06)^2/0.97^2)$. Observe that in this example, the mean of the sample means is 8.06 with a standard deviation is 0.97, and are entered into the function to determine the y-values of the graph. The 80 is a scaling factor designed to make the plot of the histogram and the normal curve coincide. Other scaling factors can be explored. You may choose to replace the mean of 8.06 and the standard deviation of 0.97 with the mean and standard deviation from your random sample. The left and right parentheses

are located on row 6. Press 2nd $\pi$, $\pi$ is located on the 5th row, right column above the ^ key. Press 2nd $e$, $e$ is located on the 8th row, left column above the LN key. Be sure to use the grey negation key when you enter $(-1/2)$. The function is shown in Figure 9.10.

6. Set the Window viewing variables in order to view the graph.

   Press WINDOW, row 1, column 2. Set Xmin to 4.5, Xmax to 12.5; Xscl to 1; Ymin to -2, being sure to use the grey negation key. Set Ymax to 300; Yscl to 1; Xres to 1. These settings are illustrated in Figure 9.11

7. View the graph.

   Press GRAPH, to view the graph, as shown in Figure 9.12.

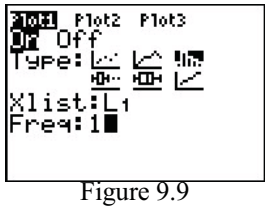


Figure 9.9

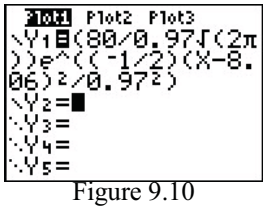


Figure 9.10

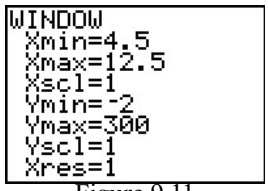


Figure 9.11

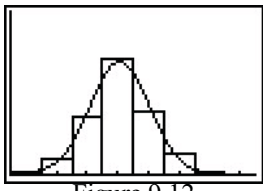


Figure 9.12

8. Turn off all plots and return the graph window to standard viewing.

   Press 2nd STAT PLOT, selecting PlotsOff and press ENTER. Press ZOOM and select 6: ZStandard to restore the default graph window settings.

9. Clear the function.

   Press Y=, and Press Y= and press CLEAR to remove all functions For each line that is not blank, place the cursor on the function and press CLEAR Press 2nd QUIT.

   Figure 9.12 indirectly shows how a sampling distribution provides information about the accuracy of a sample statistic. In this example, we learned that with a sample size of n =25 the weight losses are approximately normal. From the Empirical Rule, we know the following facts about possible sample means in this situtation, based on intervals extending 1, 2. and 3 standard deviations from the mean of 8:

   a. There is a 68%chance that the sample mean will be between 7and 9.

   b. There is a 95%chance that the sample mean will be between 6 and 10.

c. It is almost certain that the sample mean will be between 5 and 11.

## 9.3 Areas and Probabilities for Student's t-Distribution

Because Student's t-distribution differs for each possible df value, we can't summarize the probability areas in one table like we could for the standard normal distribution. We would need a separate table for each possible df value. Instead, tables for the t-distribution are tailored to specific uses.
Many calculators and computer software programs provide probabilities (areas) for specified (t-values and t-values for specified areas. For example, the TI-83 and TI-84 calculate the Student's t-distribution probability between a *lowerbound* and an *upperbound* for a specified degrees of freedom, *df*. In other words, it provides P(t> k).

**Example 9.7 Standardized Mean Weights** In Section 9.3, we considered four hypothetical samples of n = 25 people who were trying to lose weight at a clinic. Weplayed the role of the all-knowing sage and assumed we knew that $\mu = 8$ and $\sigma = 5$. If the value for $\mu$ is correct, then the standardized statistic

$$t = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\overline{x}-8}{\frac{s}{\sqrt{25}}}$$

has a t-distribution with $df = 25 - 1 = 24$. If we were to generate thousands of random samples of size 25 and draw a histogram of the resulting standardized t-statistics, they would adhere to this t-distribution.
In practice, we do not draw thousands of samples and we do not know $\mu$. Suppose we speculated that $\mu = 8$ pounds and drew one random sample, the first one given in Table 9.1 of the text, for which $\overline{x}$ = 8.32 pounds and $s = 4.74$ pounds. Are the sample results consistent with the speculation that $\mu = 8$ pounds? In other words. is a sample mean of 8.32 pounds reasonable to expect if $\mu = 8$ pounds?
The standardized statistic is

$$t = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\overline{x}-8}{\frac{s}{\sqrt{25}}} = \frac{8.32-8}{\frac{4.74}{\sqrt{25}}} = 0.34$$

**Follow these steps** to find the probability of observing a test statistic of $t = 0.34$, or greater.

1. Calculate the probability of $t = 0.34$ or greater.

   Press [2nd] [DISTR], located on row 4, column 4, above [VARS] ,to obtain the distribution function menu.

   a. Use the down arrow key, [▼], selecting 5: tcdf( , the Student's-t cumulative distribution probability function, as shown in Figure 9.13. Press [ENTER]. Enter the lowerbound, upperbound, and degrees of freedom, *df*. Type 0.34[,]100[,]24[)], as shown in Figure 9.14. Press [ENTER] to execute the command. The output from the TI calculator is displayed in Figure

9.15.

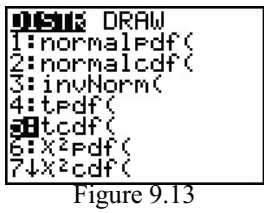


Figure 9.13

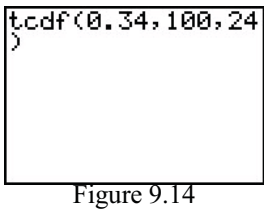


Figure 9.14

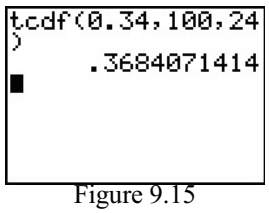


Figure 9.15

This statistic,$t = 0.34$, tells us that the sample mean of 8.32 is only about 0.3684 of a standard error above 8, which is certainly consistent with a population mean weight loss of 8 pounds.

**Variations in Finding Areas for a Student's t-distribution**

Variation 1: **Follow these steps** to find the probability of observing a test statistic of $t = 0.34$, or less, given that the degrees of freedom is 24.

1. Calculate the probability of $t = 0.34$ or less.

   Press 2nd DISTR, located on row 4, column 4, above VARS ,to obtain the distribution function menu.

   a. Use the down arrow key, ▼, selecting 5: tcdf( , the Student's-t cumulative distribution probability function, as shown in Figure 9.16. Press ENTER. Enter the lowerbound, upperbound, and degrees of freedom, $df$. Type 0.34 , 100 , 24 ), as shown in Figure 9.14. Press ENTER to execute the command. The output from the TI calculator is displayed in Figure 9.15.

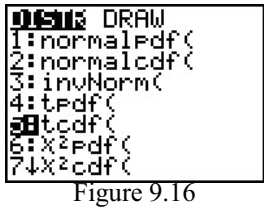


Figure 9.16

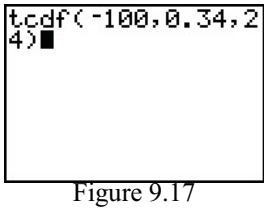


Figure 9.17

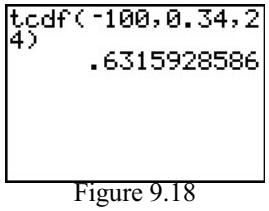


Figure 9.18

This TI calculator output tells us that the probability of finding a value of $t = 0.34$ or less, given that the degrees of freedom is 24, is about 0.6316 or 63.16% of the time.

Variation 2: **Follow these steps** to find the probability of observing a test statistic between $t = -1.17$ and $t = +2.27$, given that the degrees of freedom is 9.

1. Calculate the probability of observing a value of $t$ between $t = -1.17$ and $t = +2.27$.

Press 2nd DISTR, located on row 4, column 4, above VARS ,to obtain the distribution function menu.

a. Use the down arrow key, ▼, selecting 5: tcdf( , the Student's-t cumulative distribution probability function, as shown in Figure 9.19. Press ENTER. Enter the lowerbound, upperbound, and degrees of freedom, $df$. Type -1.17 , 2.27 , 9 ), as shown in Figure 9.20. Press ENTER to execute the command. The output from the TI calculator is displayed in Figure 9.21.

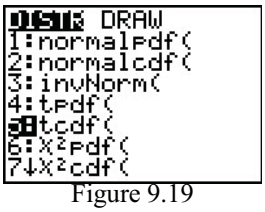


Figure 9.19

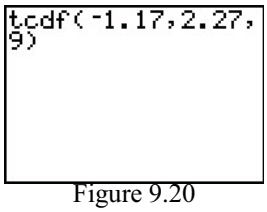


Figure 9.20

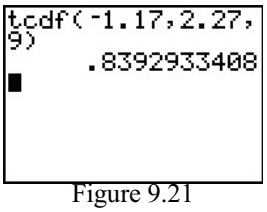


Figure 9.21

This TI calculator output tells us that the probability of finding a value of $t$ between $t = -1.17$ and $t = +2.27$, given that the degrees of freedom is 9 is about 0.8393 or 83.93% of the time.

# Chapter 10
# Estimating Proportions With Confidence

In this Chapter we will learn about confidence intervals, including what they are, how to interpret them and the generic fonnat to use to construct one for the five parameters covered in Chapter 9. We will also learn the specifics about constructing and interpreting a confidence interval for one population proportion and for the difference in two population proportions.
Confidence intervals involving one population proportion,$p$, and the difference in two population proportions for independent samples, $p_1 - p_2$ are covered in this chapter.

A **confidence interval** or **interval estimate** for any of the five parameters can be expressed as:

Sample estimate $\pm$ :multiplier $\times$ standard error

The **multiplier** is a number based on the confidence level desired, and determined from the standard normal distribution (for proportions) or Student's t-distribution (for means).
Details are provided in the individual modules.

After reading this chapter you should be able to:

1. Find the appropriate $z$-multiplier for a specified level of confidence.
2. Find a confidence interval for a proportion.
3. Compute a confidence interval for the difference in two proportions using summarized data.

**Keystrokes Introduced**

1. [2nd] [DISTR], selecting 3: invNorm(. The arguments are 3: invNorm(*area[ $\mu$,$\sigma$])*. The command computes the inverse cumulative normal distribution function for a given area under the normal distribution curve specified by $\mu$ and $\sigma$. If $\mu$ and $\sigma$ are not specified, the default is the standard normal distribution.
2. [2nd] [DISTR] [►] accessing the DRAW menu, selecting 1: ShadeNorm(. The arguments are 1: ShadeNorm(*lowerbound,upperbound[ $\mu$,$\sigma$])*.
   The command draws the normal density function specified by $\mu$ and $\sigma$ and shades the area between *lowerbound* and *upperbound*. If $\mu$ and $\sigma$ are not specified, the default is the standard normal distribution.
3. [STAT] [►] [►], accessing the TESTS menu, selecting B: 2-PropZInt. **2-PropZInt** computes a two-proportion $z$ confidence interval. The arguments in the two-proportion $z$ confidence interval are B: 2-PropZInt ($x_1$,$n_1$,$x_2$,$n_2$[,*confidence level]*.

## 10.1 Finding the Multiplier $z^*$

The multiplier is denoted by $z^*$ and is found using the standard normal distribution. Values of the multiplier for the most common confidence levels used by researchers are shown in Table 10.1 of the text and can be found using the TI-83 SE or TI-84 SE calculator.

**Example: Finding Appropriate z-Multipliers**
**For a Specified Level of Confidence (90%, 95%, 98%, 99%)**

**Follow these steps** to find an appropriate $z$-multiplier for a specified level of confidence.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Find the appropriate $z^*$-multiplier for a specified level of confidence.
   a. In order to find the appropriate $z$-multiplier for a **90% level of confidence**, we use the invNorm function requiring the area to the left of the $z$-multiplier. Therefore for a 90% level of confidence, you will have 10% outside the interval, with 5% (0.05) in each tail of the standard normal curve, as shown in Figure 10.1. The area to the left of the positive $z$-multiplier will then be $0.90 + 0.05 = 0.95$. Take the following steps:
   On the homescreen, press [2nd] [DISTR], located on the fourth row, column four, above [VARS]. Select 3: invNorm( .Type 0.95 [)] [ENTER]. The results are shown in Figure 10.2, indicating the appropriate $z$-multiplier for a 90% level of confidence is 1.645, rounded to 3 decimal places or 1.65, rounded to 2 decimal places.

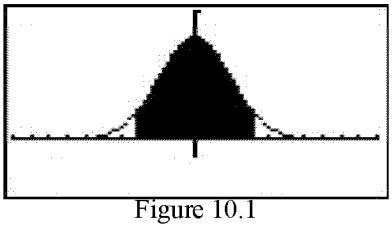


Figure 10.1

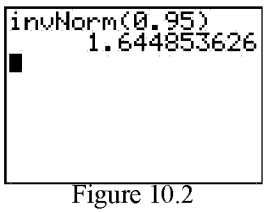



Figure 10.2

   b. In order to find the appropriate $z$-multiplier for a **95% level of confidence**, we use the invNorm function requiring the area to the left of the

$z$-multiplier. Therefore for a 95% level of confidence, you will have 5% outside the interval, with 2.5% (0.025) in each tail of the standard normal curve, as shown in Figure 10.1. The area to the left of the positive $z$-multiplier will then be $0.95 + 0.025 = 0.975$. Take the following steps:
On the homescreen, press 2nd DISTR, located on the fourth row, column four, above VARS. Select 3: invNorm( .Type 0.975 ) ENTER. The results are shown in Figure 10.3, indicating the appropriate $z$-multiplier for a 95% level of confidence is 1.96, rounded to 2 decimal places.
To create the graphic illustrating the 95% confidence interval under the standard normal curve, take the following steps:
On the homescreen, press 2nd DISTR ▶, selecting DRAW and 1: ShadeNorm( as shown in Figure 10.4. Type -1.96,1.96 ) ENTER. The arguments of the 1: ShadeNorm( function are *lowerbound, upperbound.* The standard normal curve with the 95% shaded confidence interval is shown in Figure 10.5.

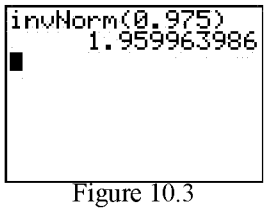


Figure 10.3

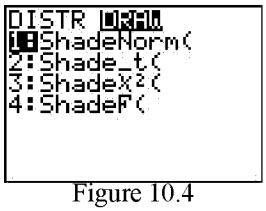


Figure 10.4

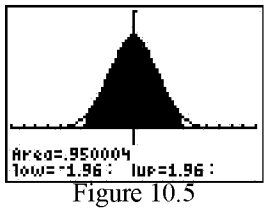


Figure 10.5

c. In order to find the appropriate $z$-multiplier for a **98% level of confidence**, we use the invNorm function requiring the area to the left of the $z$-multiplier. Therefore for a 98% level of confidence, you will have 2% outside the interval, with 1% (0.01) in each tail of the standard normal curve, as shown in Figure 10.1. The area to the left of the positive $z$-multiplier will then be $0.98 + 0.01 = 0.99$. Take the following steps:
On the homescreen, press 2nd DISTR, located on the fourth row, column four, above VARS. Select 3: invNorm( .Type 0.99 ) ENTER. The results are shown in Figure 10.6, indicating the appropriate $z$-multiplier for a 98% level of confidence is 2.33, rounded to 2 decimal places.
To create the graphic illustrating the 98% confidence interval under the standard normal curve, take the following steps:
On the homescreen, press 2nd DISTR ▶, selecting DRAW and 1: ShadeNorm( as shown in Figure 10.7. Type -2.33,2.33 ) ENTER. The arguments of the 1: ShadeNorm( function are *lowerbound, upperbound.* The standard normal curve with the 98% shaded confidence interval is shown

in Figure 10.8.

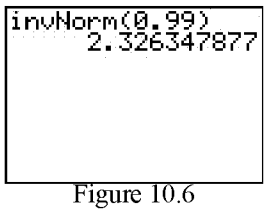


Figure 10.6

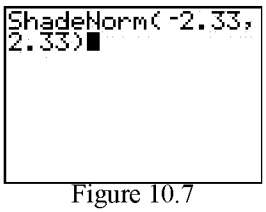


Figure 10.7

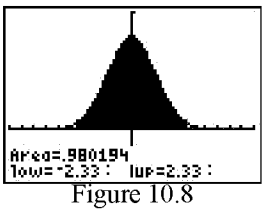


Figure 10.8

d. In order to find the appropriate $z$-multiplier for a **99% level of confidence**, we use the invNorm function requiring the area to the left of the $z$-multiplier. Therefore for a 99% level of confidence, you will have 1% outside the interval, with 0.5% (0.005) in each tail of the standard normal curve, as shown in Figure 10.1. The area to the left of the positive $z$-multiplier will then be $0.99 + 0.005 = 0.995$. Take the following steps:
On the homescreen, press 2nd DISTR, located on the fourth row, column four, above VARS. Select 3: invNorm( .Type 0.995 ) ENTER. The results are shown in Figure 10.9, indicating the appropriate $z$-multiplier for a 99% level of confidence is 2.58, rounded to 2 decimal places.
To create the graphic illustrating the 99% confidence interval under the standard normal curve, take the following steps:
On the homescreen, press 2nd DISTR ▶, selecting DRAW and 1: ShadeNorm( as shown in Figure 10.10. Type -2.58,2.58 ) ENTER. The arguments of the 1: ShadeNorm( function are *lowerbound, upperbound*. The standard normal curve with the 99% shaded confidence interval is shown in Figure 10.11.

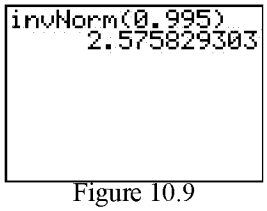


Figure 10.9

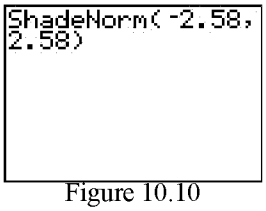


Figure 10.10

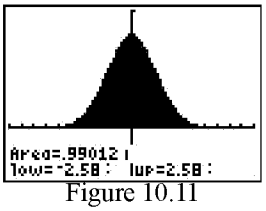


Figure 10.11

**Example 10.3 - Is There Intelligent Life on Other Planets**

In a 1997 Marist Institute survey of 935 randomly selected Americans, 60% (60% of 935 = 561) of the sample answered "yes" to the question "Do you think there is intelligent life on other planets?" (Source: www.mipo.marist.edu). Let's use this sample estimate to calculate a 90% confidence interval for the proportion of an Americans who believe there is intelligent life on other planets.

**Follow these steps** to find a confidence for a proportion..

1. Preparations:
   a. Turn off all"Y=" functions.

Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

b. Clear all lists in the Stat editor.

Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Find the a confidence interval for the proportion of all Americans who believe there is intelligent life on other planets.

a. In order to find a **90% confidence interval** for the proportion of all Americans who believe there is intelligent life on other planets, select [STAT] [►] [►] accessing the TESTS menu. Press [▼] several times to select A: 1-PropZInt, as shown in Figure 10.12. Press [ENTER]. Type 561 in the $x$: input area, pressing [ENTER] after the entry. Type 935 in the $n$: input area, pressing [ENTER] after the entry. Type 0.90 in the C-Level: input area, pressing [ENTER] after the entry. Highlight calculate and press the [ENTER] key. The screen is shown in Figure 10.13. The results of executing the command are shown in Figure 10.14.

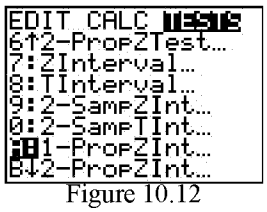


Figure 10.12

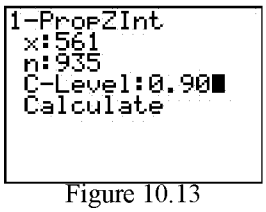


Figure 10.13

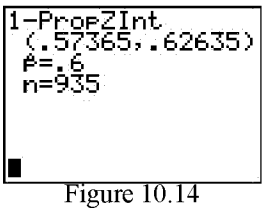


Figure 10.14

With 90% confidence, we can say that in 1997 the proportion of Americans who believed there is intelligent life on other planets was in the interval from 0.574 to 0.626 (0.60±0.026).

If s.e.($\widehat{p}$) is calculated as $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{m}} = \sqrt{\frac{.6(1-.6)}{935}} = 0.016$, then the confidence interval can be calculated manually. The appropriate $z$-multiplier for a 90% level of confidence is 1.65, rounded to 2 decimal places. In the traditional format, the 90% confidence interval might be organized in this format.

$$\widehat{p} \pm z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$
$$0.60 \pm 1.65 \times 0.16$$
$$0.574 \text{ to } 0.626$$

b. We can increase our confidence that the interval covers the truth by increasing the confidence level. We pay a price, because the multiplier will be larger and the interval will be wider. For instance, we could find a 95% confidence interval. In order to find a **95% confidence interval** for the proportion of all Americans who believe there is intelligent life on other planets, select [STAT] [►] [►] accessing the TESTS menu. Press [▼]

several times to select A: 1-PropZInt.., as shown in Figure 10.12. Press ENTER. Type 561 in the x: input area, pressing ENTER after the entry. Type 935 in the n: input area, pressing ENTER after the entry. Type 0.95 in the C-Level: input area, pressing ENTER after the entry. Highlight calculate and press the ENTER key. The screen is shown in Figure 10.15. The results of executing the command are shown in Figure 10.16.

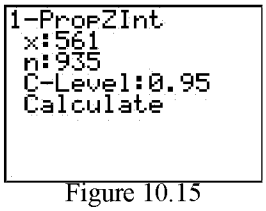



Figure 10.15

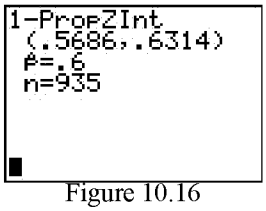



Figure 10.16

With 95% confidence, we can say that in 1997 the proportion of Americans who believed there is intelligent life on other planets was in the interval from 0.569 to 0.631 ($0.60 \pm 0.31$).

If s.e.($\widehat{p}$) is calculated as $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{m}} = \sqrt{\frac{.6(1-.6)}{935}} = 0.016$, then the confidence interval can be calculated manually. The appropriate $z$-multiplier for a 95% level of confidence is 1.96, rounded to 2 decimal places. In the traditional format, the 95% confidence interval might be organized in this format.

$$\widehat{p} \pm z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$
$$0.60 \pm 1.96 \times 0.16$$
$$0.569 \text{ to } 0.631$$

c. In order to find a **98% confidence interval** for the proportion of all Americans who believe there is intelligent life on other planets, select STAT ► ► accessing the TESTS menu. Press ▼ several times to select A: 1-PropZInt.., as shown in Figure 10.12. Press ENTER. Type 561 in the x: input area, pressing ENTER after the entry. Type 935 in the n: input area, pressing ENTER after the entry. Type 0.98 in the C-Level: input area, pressing ENTER after the entry. Highlight calculate and press the ENTER key. The screen is shown in Figure 10.15. The results of executing the command are shown in Figure 10.16.

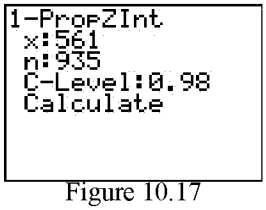



Figure 10.17

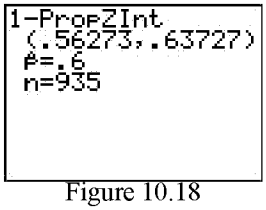



Figure 10.18

With 98% confidence, we can say that in 1997 the proportion of Americans who believed there is intelligent life on other planets was in the interval from 0.563 to 0.637 ($0.60 \pm 0.37$).

If s.e.($\widehat{p}$) is calculated as $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{m}} = \sqrt{\frac{.6(1-.6)}{935}} = 0.016$, then the con-

fidence interval can be calculated manually. The appropriate $z$-multiplier for a 98% level of confidence is 2.33, rounded to 2 decimal places. In the traditional format, the 98% confidence interval might be organized in this format.

$$\widehat{p} \pm z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$
$$0.60 \pm 2.33 \times 0.16$$
$$0.563 \text{ to } 0.637$$

## 10.2 CI for the Difference in Two Proportions

The objective in this section is to form a confidence interval for the difference between two population proportions. Again,we start with the same general format that we introduced in Module 0 for the five parameters of interest,

Sample estimate $\pm$ :multiplier $\times$ standard error

**Conditions for a Confidence Interval**
**for the Difference in TwoProportions**
As is the case for all statistical inference procedures, there are conditions that should be present in order to use the confidence interval just described. For a confidence interval for the difference between two proportions, the principal conditions have to do with the sample sizes for the observed samples. Both conditions must hold.

Condition 1: Sample proportions are available based on independent, randomly selected samples from the two populations.

Condition 2: All of the quantities $n_1\widehat{p}_1$, $n_1(1-\widehat{p}_1)$, $n_2\widehat{p}_2$, and $n_2(1-\widehat{p}_2)$are at least 10. These quantities represent the counts observed in the category of interest and not in that category, respectively, for the two samples.

Notice that these are the same conditions required for the sampling distribution of $\widehat{p}_1 - \widehat{p}_2$ to be approximately normal, except that the sample size condition now applies using $\widehat{p}$'s instead of the $p$'s.

**Example 10.10 - Snoring and Heart Attacks**

P.G. Norton and E. V.Dunn conducted a study to determine whether there is a relationship between snoring and risk of heart disease (Norton and Dunn, 1985). They found that of the 1105 snorers they sampled, 86 had heart disease, while only 24 of the 1379 nonsnorers had heart disease.We will define population 1 to be snorers and population 2 to be nonsnorers. In each case, $p$ is the (unknown) population proportion with heart disease.
How much difference is there between the proportions with heart disease for the two populations? We can answer this question with a confidence interval. First recognize that the conditions for calculating this confidence interval are satisfied. Condition 1 is satisfied because the researchers collected independent samples

from each population, and Condition 2 is satisfied because the relevant observed counts are all greater than 10. In the group that snores, the numbers with and without heart disease, respectively, are 86 and 1019, while in the group that doesn't snore, the numbers with and without heart disease are 24 and 1355, respectively.

**Follow these steps** to find a confidence interval for the difference in two proportions

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Find **Follow these steps** to find confidence intervals for the difference in two proportions.

   a. Find the 90% confidence interval for the difference in two proportions. Select [STAT] [►] [►], accessing the TESTS menu. Select B: 2-PropZInt, as shown in Figure 10.19, and press [ENTER]. Use the down arrow key, [▼], to enter 86 in $x_1$:. Use the down arrow key, [▼], to enter 1105 in $n_1$:. Use the down arrow key, [▼], to enter 24 in $x_2$:. Use the down arrow key, [▼], to enter 1379 in $n_2$:. Use the down arrow key, [▼], and enter the confidence level 0.90. Use the down arrow key, highlighting Calculate and [ENTER] to execute the command, as shown in Figure 10.20. The results are shown in Figure 10.21.

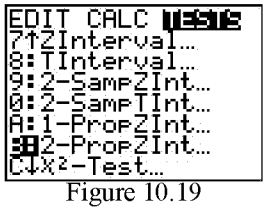


Figure 10.19

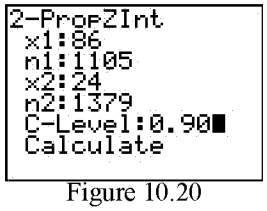


Figure 10.20

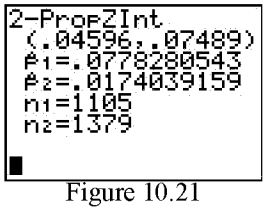


Figure 10.21

   The TI output, shown in Figure 10.21, indicates the 90% confidence interval 0.046 to 0.075.
   b. Find the 95% confidence interval for the difference in two proportions. Select [STAT] [►] [►], accessing the TESTS menu. Select B: 2-PropZInt,

as shown in Figure 10.22, and press ENTER. Use the down arrow key, ▼, to enter 86 in $x_1$:. Use the down arrow key, ▼, to enter 1105 in $n_1$:. Use the down arrow key, ▼, to enter 24 in $x_2$:. Use the down arrow key, ▼, to enter 1379 in $n_2$:. Use the down arrow key, ▼, and enter the confidence level 0.95. Use the down arrow key, highlighting Calculate and ENTER to execute the command, as shown in Figure 10.23. The results are shown in Figure 10.24.

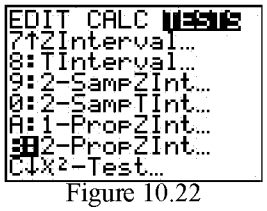


Figure 10.22

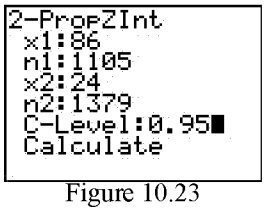


Figure 10.23

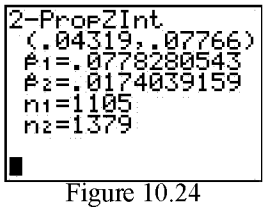


Figure 10.24

The TI output, shown in Figure 10.24, indicates the 95% confidence interval 0.043 to 0.078

c. Find the 99% confidence interval for the difference in two proportions. Select STAT ► ►, accessing the TESTS menu. Select B: 2-PropZInt, as shown in Figure 10.25, and press ENTER. Use the down arrow key, ▼, to enter 86 in $x_1$:. Use the down arrow key, ▼, to enter 1105 in $n_1$:. Use the down arrow key, ▼, to enter 24 in $x_2$:. Use the down arrow key, ▼, to enter 1379 in $n_2$:. Use the down arrow key, ▼, and enter the confidence level 0.99. Use the down arrow key, highlighting Calculate and ENTER to execute the command, as shown in Figure 10.26. The results are shown in Figure 10.27.

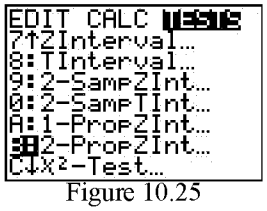


Figure 10.25

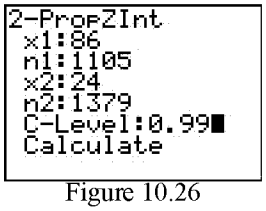


Figure 10.26

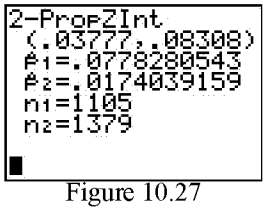


Figure 10.27

The TI output, shown in Figure 10.27, indicates the 99% confidence interval 0.038 to 0.083.

As is always the case, the higher the level of confidence, the wider the interval. From these Intervals, It appears that the proportion of snorers with heart disease in the population is about 4% to 8% higher than the proportion of nonsnorers with heart disease.

# Chapter 11
# Estimating Means with Confidence

In Chapter 10 we learned how to form and interpret a confidence interval for a population proportion and for the difference in two population proportions. Remember that a confidence interval is an interval of values computed from sample data that is likely to include the true population value. For example, based on a sample survey of 1000 Americans, we may be able to say with 95% confidence that the interval from .23 to .29 (or 23% to 29%) will cover the true proportion of all Americans who favor the legalization of marijuana. The confidence level for an interval is a proportion or percent, such as 95%, that describes our confidence in theprocedure we used to determine the interval. Weare confident that in the long run, the stated confidence level is the percent of the confidence intervals computed with the procedure that will actually. contain the true population value.
In this chapter, we expand the use of confidence intervals to parameters involving means of populations, including the population mean of one quantitative variable, the population mean of paired differences, and the difference in means for two populations. We may, for example, wish to know the mean number of hours per day that college students watch television. As an example of the mean of paired differences, we may want to know how much the mean IQ would change if everyone in the population were to listen to classical music for 30 minutes. As an example of the difference in two population means, we may want to know the difference in effectiveness of two diets, based on comparing the mean weight that would be lost if everyone in a population were to go on one diet versus if they were to go on the other diet. In each of these types of situations, the objective is the same. We use sample information to form a confidence interval estimate of a population value.

After reading this chapter you should be able to:

1. Find a t-multiplier for a given level of confidence and a given degrees of freedom.
2. Check the conditions before finding a confidence interval for a mean.
3. Construct a confidence interval for a single mean, using raw data.
4. Construct a confidence interval for a single mean, using summarized data.
5. Find paired differences for raw data.
    a. Check the conditions before finding a confidence interval for the mean of paired data.
    b. Construct a confidence interval for the mean of paired data, using raw data.
6. Check the conditions before finding a confidence interval for the difference between two independent means.

7. Construct a confidence interval for the difference between two independent means - both pooled and unpooled versions.

**Keystrokes Introduced**

1. **solve(***expression,variable,guess* [,{*lower,upper* }]) The function **solve(** is available only from CATALOG or from within a program. It returns a solution (root) of *expression* for *variable*, given an initial *guess*, and *lower* and *upper* bounds with which the solution is sought. The default for *lower* is $-1E99$. The default for *upper* is $1E99$. **Solve(** is valid for only real numbers.
2. [2nd] [ENTRY] recalls the previous expression.
3. [STAT] [►] [►], accessing the TESTS menu and selecting 8: TInterval. **TInterval** computes a $t$ confidence interval. The arguments are 8: TInterval (*listname, freqlist,confidence level).*
4. [STAT] [►] [►], accessing the TESTS menu and selecting 0: 2-SampTInt... **2-SampTInt...** computes a two-sample $t$ confidence interval. *Pooled* = Yes polls variances; *Pooled* = No does not pool variances. The arguments are 2-SampTInt... (*listname1,listname2, freqlist1,freqlist2,, confidence level,pooled).*

## 11.1 Finding the t-multiplier

**Example 11.4 Finding the t* Values**
Suppose you want to find a 95% confidence interval and also a 99% confidence interval for the mean of a population based on a sample of $n = 25$ values. For the 95% interval, the appropriate multiplier is the value $t^*$ such that the area between $-t^*$ and $t^*$ is 0.95 or 95%, for a Student's $t$-distribution with degrees of freedom $n - 1 = 25 - 1 = 24$.

**Follow these steps** to find a t-multiplier for a given level of confidence and a given degrees of freedom.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Solve for the $t$-multiplier.
   a. Place the solve( function on the homescreen.

      On the homescreen, press [2nd] [CATALOG]; [CATALOG] is located on the bottom row, column two, over [0]. Press [T], located over the [4], and the up arrow [▲], a number of times, locating solve(, as shown in Figure 11.1. Press [ENTER] to place the function on the homescreen.


      Figure 11.1

   b. Recall that the function **solve(***expression*,*variable*,*guess*[,{*lower*,*upper*}]) requires the *expression*,*variable*,*guess*[,{*lower*,*upper*}]).
      (i) Enter the *expression*, $tcdf(X, 1E99, 24) - 0.025$, that will used to solve for the $t^*$ multiplier.

      Press [2nd] [DISTR], selecting 5: tcdf(, and press [ENTER]. Enter the *lowerbound* by pressing [$X,T,\theta,n$]. Enter the *upperbound* by pressing [1] [EE] [9] [9]; [EE] is located on the sixth row, column two, above the [,]. Enter the degrees of freedom, $df$, by typing [9] [)]. Enter the area in each tail of the confidence interval: $\frac{(1-0.95)}{2} = 0.025$, by typing $-0.025$ and [,].

      (ii) Enter the *variable*, $X$, to be solved for by pressing [$X,T,\theta,n$] and [,].

      (iii) Enter a *guess*, 3 by pressing [3]. Other guesses will also work. Close the function by entering [)]. Press [ENTER] to execute the command. The expression and the evaluation are shown in Figure 11.2.

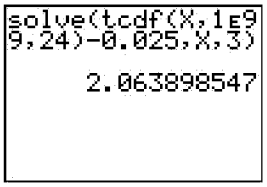



      Figure 11.2

The value of the $t$-multiplier for the 95% confidence interval, with 24 degrees

of freedom, is 2.06 and agrees with the value found in Table A.2 of the text.

c. Find the value of the $t$-multiplier for the 99% confidence interval.

Press [2nd] [ENTRY]; [ENTRY] is located above the [ENTER] key, displaying the previous expression. Enter the area in each tail of the confidence interval: $\frac{(1-0.99)}{2} = 0.005$, by typing $-0.005$. Use the up arrow, [▲], placing the cursor on 0.025. Change the 0.025 to 0.005, as shown in Figure 11.3. Press [ENTER] to execute the command. The expression and the evaluation are shown in Figure 11.4.

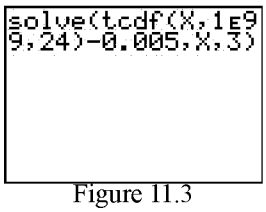


Figure 11.3

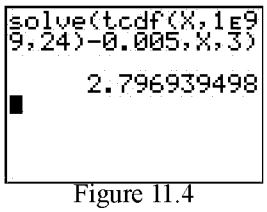


Figure 11.4

The value of the $t$-multiplier for the 99% confidence interval, with 24 degrees of freedom, is 2.80 and agrees with the value found in Table A.2 of the text.

## 11.2 Confidence Intervals for One Mean

Researchers often are interested in a very simple question about a population: if we were to measure a quantitative variable for every unit in the population, what would be the mean or average of those values? For example, what is the mean age of death for the population of left-handed adults? What is the mean human body temperature for healthy adults? What is the mean per capita income for all adults in your state? What is the mean price of a gallon of gasoline last week in the United States? What is the mean amount of sleep college students get per night?
All of these questions can be answered by taking an appropriate sample from the population of interest, and computing **a confidence interval estimate of the mean** of the population, which is an interval of values computed from the sample data that we can be fairly confident covers the true population mean. How certain we can be is determined by the confidence level we choose to use.

### Conditions Required for Using the $t$ Confidence Interval

The methods presented in this section are derived mathematically by assuming that the sample has been randomly selected from a population in which the response variable has a normal distribution. The t-interval procedure, however, is a robust procedure because it works well over a wide range of situations. Put another way, the stated confidence level for a t-interval is approximately correct in many situations where the assumption about a normally distributed response variable is not

correct.
For large sample sizes, a t-interval is a valid estimate of the population mean even in the presence of skewness. For small sample sizes. a t-interval can be used if the data are not skewed and contain no outliers.

**Two Situations for Which a t Confidence Interval for One Mean Is Valid**

Situation 1: The population of the measurements is bell-shaped and a random sample of any size is measured. In practice, for small samples, the data should show no extreme skewness and should not contain any outliers.

Situation 2: The population of measurements is not bell-shaped, but a large random sample is measured. A somewhat arbitrary definition of a "large" sample is $n > 30$, but if there are extreme outliers, it is better to have a larger sample.

Before calculating a confidence interval for a mean, first check that one of the situations just described holds, To look for outliers or skewness,plot the data using either a histogram, boxplot. dotplot, or stemplot.

**Example 11.5 - Are Your Sleeves Too Short? The Mean Forearm Length of Men**

People are always interested in comparing themselves to others. If you are male, how does the length of your forearm, from elbow to wrist, compare to the average for the population of men? Suppose that the forearm lengths (cm) for a randomly selected sample of $n = 9$ men are

25.2 24.0 26.5 25.5 28.0 27.0 23.0 25.0 25.0

Table 11.1

**Follow these steps** to construct a histogram checking the conditions required for using the t confidence interval.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

a. Enter the data found in Table 11.1.

Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the data: 25.5, 24.0, 26.5, ... pressing ENTER after each entry, as shown in Figure 11.5.

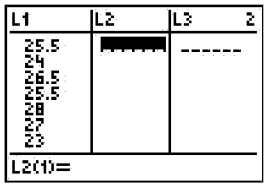



Figure 11.5

b. Plot the statistical data by creating a histogram of the forearm lengths (cm) of the randomly selected sample of $n = 9$ men .

Press 2nd STAT PLOT accessing the StatPlot menu, as shown in Figure 11.6. Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram. Press ENTER. Use the down arrow key to select L1 as the list, 2nd L1, as shown in Figure 11.7. Enter a frequency of 1 in Freq:. Press ZOOM 9: ZoomStat to view the histogram as shown in Figure 11.8.

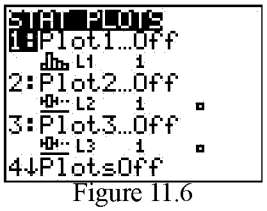



Figure 11.6

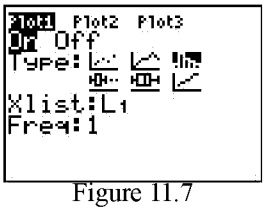



Figure 11.7

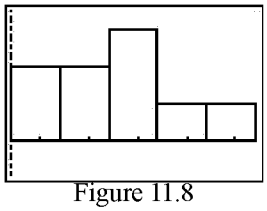


Figure 11.8

Figure 11.8 shows a histogram of these values. There is no obvious skewness and there are no outliers. Assuming the sample can be thought of as a random sample, Situtation 1 holds and the necessary conditions for computing a confidence interval for the mean are present.

**Follow these steps** to construct a confidence interval for a single mean.

1. Select STAT ► ►, accessing the TESTS menu. Select 8: TInterval, as shown in Figure 11.9, and press ENTER. Highlight Data and press ENTER to make it "stick". Use the down arrow key, ▼, to enter 2nd L1 in List:. Enter a frequency of 1 in Freq:. Type 0.95 in the confidence interval C-Level:. Use the down arrow key, highlighting Calculate and ENTER to execute the

command, as shown in Figure 11.10. The results are shown in Figure 11.11.

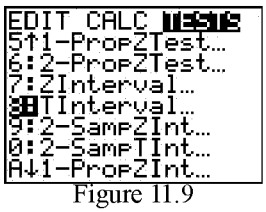


Figure 11.9

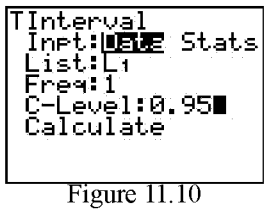


Figure 11.10

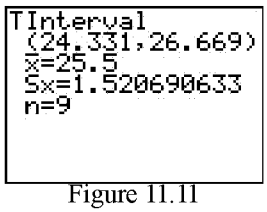


Figure 11.11

The TI output, shown in Figure 11.11, indicates the confidence interval (24.331,26.669). The sample mean, $\overline{x}$, is $25.5$ and the sample standard deviation, $s$, is 1.52, rounded to 2 decimal places. We can say with $95\%$ confidence that in the population represented by the sample, the mean forearm length is between 24.33 and 26.67 cm.

## Computing a Confidence Interval For a Single Mean

Before calculating a confidence interval for a mean, first check that one of the situations just described holds, To look for outliers or skewness,plot the data using either a histogram, boxplot. dotplot, or stemplot.

**Example 11.7 - What Type of Students Sleep More?**

On the second day of classes in Spring 2000, a Monday, students in two statistics classes at the University of California at Davis were given a survey. One of the questions asked was, "How many hours of sleep did you get last night, to the nearest hour?" One class was Statistics 10, a statistical literacy course for liberal arts majors ($n = 25$), while the other was Statistics 13, a large introductory class in statistical methods for more technical majors ($n = 148$). Here is the Minitab output that summarizes the information on this question for the two classes:

| class | N | Mean | StDev | SE Mean |
|---|---|---|---|---|
| Stat 10 | 25 | 7.66 | 1.34 | 0.27 |
| Stat 13 | 148 | 6.81 | 1.73 | 0.14 |

Table 11.2

Checking the Conditions
For each group, we must make the assumption that these students are equivalent to a random sample from the population of interest. Although not strictly true, the students in these two classes are probably representative of all students who take these and similar classes over the years. Second, for the Statistics 10 class, the sample size is only 25, so we must assume that the population of sleep times is approximately bell-shaped. A histogram of the sample values (not shown) indicated that this assumption is reasonable.

**Follow these steps** to construct a confidence interval for each of the means.

1. Construct a confidence interval for the mean of Statistics 10.

Select STAT ► ►, accessing the TESTS menu. Select 8: TInterval, as shown in Figure 11.12, and press ENTER. Highlight Stats and press ENTER to make it "stick". Use the down arrow key, ▼, to enter 7.66 in $\overline{x}$:. Use the down arrow key, ▼, to enter 1.34 in $S_x$:. Use the down arrow key, ▼, to enter 25 in $n$:. Use the down arrow key, ▼, to type 0.95 in the confidence interval C-Level:. Use the down arrow key, highlighting Calculate and ENTER to execute the command, as shown in Figure 11.13. The results are shown in Figure 11.14.

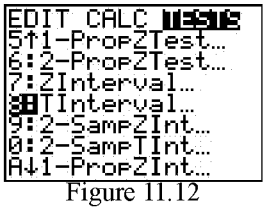


Figure 11.12

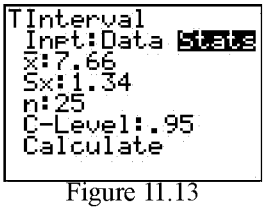


Figure 11.13

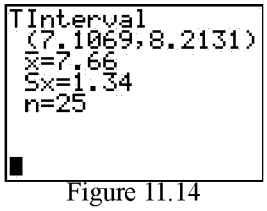


Figure 11.14

The TI output, shown in Figure 11.14, indicates the confidence interval (7.1069,8.2131). The sample mean, $\overline{x}$, is 7.66 and the sample standard deviation, $s$, is 1.34. We can say with 95% confidence that ithe average hours of sleep for all students similar to those in Statistics 10 is covered by the interval from 7.10 to 8.21 hours.

1. Construct a confidence interval for the mean of Statistics 13.

Select STAT ► ►, accessing the TESTS menu. Select 8: TInterval, as shown in Figure 11.12, and press ENTER. Highlight Stats and press ENTER to make it "stick". Use the down arrow key, ▼, to enter 6.81 in $\overline{x}$:. Use the down arrow key, ▼, to enter 1.73 in $S_x$:. Use the down arrow key, ▼, to enter 148 in $n$:. Use the down arrow key, ▼, to type 0.95 in the confidence interval C-Level:. Use the down arrow key, highlighting Calculate and ENTER to execute the command, as shown in Figure 11.16. The results are shown in Figure 11.14.

Figure 11.15

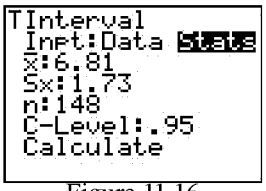


Figure 11.16

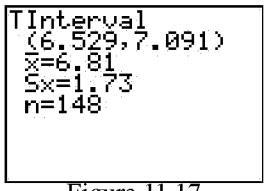


Figure 11.17

The TI output, shown in Figure 11.17, indicates the confidence interval (6.529,7.091). The sample mean, $\overline{x}$, is 6.81 and the sample standard deviation, $s$, is 1.73. We can say with 95% confidence that ithe average hours of sleep for all students similar to those in Statistics 13 is covered by the interval from 6.53 to 7.09 hours.

Therefore, it appears that the liberal arts students' average hours of sleep really is higher than the general students average hours of sleep.

## 11.3 Confidence Interval For the Mean of Paired Data

Remember that an important special case of a single mean of a population occurs when two quantitative variables are collected in pairs and we desire information about the difference between the two variables. For example. we might want to know the average difference between Verbal SAT scores for individuals before and after taking a training course. Or we may want to know the average difference in manual dexterity for the dominant and nondominant hands. In both of these examples, we could collect two measurements from each individual in the sample. To analyze paired data, we begin by calculating the difference in the two measurements for each pair in the sample.

### Checking the Conditions Before Finding a Confidence Interval for Paired Data

Once the differences have been computed, the conditions required are the same as they are for a $t$-interval for one population mean. The distinction is that the conditions must hold for the data set of differences. Let's rewrite the two situations for which the $t$-interval fonnula is appropriate, using the tenninology for the differences:

*Situation 1* : The population of differences is bell-shaped and a random sample of any size is measured. In practce, for small samples, the differences in the sample should show no extreme skewness and should not contain any outliers.
*Situation 2* : The population of differences is not bell-shaped, but a large random sample is measured. A somewhat arbitrary definition of a "large" sample is $n \geq 30$ pairs, but if there are extreme outliers in the sample of differences, it is better to have a larger sample.

Before calculating a confidence interval for the population mean of the differences, check that there are enough pairs to satisfy situation 2. If not, examine a plot of the differences - a histogram, dotplot, boxplot or stem-and-leaf plot - to make sure there are no excessive outliers or extreme skewness.

**Example 11.9 - Screen Time - Computer versus TV**

The 25 students in a liberal arts course in statistical literacy were given a survey that inc1uded questions on how many hours per week they watched television and how many hours a week they used a computer. The responses are shown in Table

12.3.

| Student | Computer | TV | Student | Computer | TV |
|---|---|---|---|---|---|
| 1 | 30 | 2.0 | 14 | 5 | 6.0 |
| 2 | 20 | 1.5 | 15 | 8 | 20.0 |
| 3 | 10 | 14.0 | 16 | 30 | 20.0 |
| 4 | 10 | 2.0 | 17 | 40 | 35.0 |
| 5 | 10 | 6.0 | 18 | 15 | 15.0 |
| 6 | 0 | 20.0 | 19 | 40 | 5.0 |
| 7 | 35 | 14.0 | 20 | 3 | 13.5 |
| 8 | 20 | 1.0 | 21 | 21 | 1.0 |
| 9 | 2 | 14.0 | 22 | 2 | 4.0 |
| 10 | 5 | 10.0 | 23 | 9 | 0.0 |
| 11 | 10 | 15.0 | 24 | 14 | 14.0 |
| 12 | 4 | 2.0 | 25 | 21 | |
| 13 | 50 | 10.0 | | | |

Table 11.3

**Follow these steps** to construct a modified boxplot checking the conditions required for using the t confidence interval.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the data found in Table 11.3.

   Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the data: 30, 20, 10 ... pressing [ENTER] after each entry, as shown in Figure 11.18.
   Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the data: 2.0, 1.5, 14.0 ... pressing [ENTER] after each entry, as shown in Figure 11.18.

b. Finding paired differences from raw data.

Place the cursor on list L3, at the top and type 2nd L1 - 2nd L2, as shown in Figure 11.19. Press ENTER to obtain the paired differences, as shown in Figure 11.20.

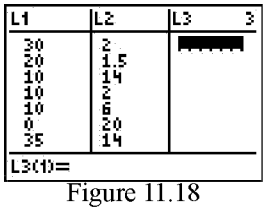


Figure 11.18

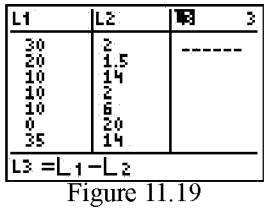


Figure 11.19

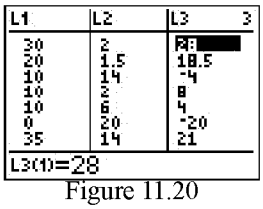


Figure 11.20

c. Checking the conditions before finding a confidence interval for paired data.
Plot the statistical data by creating a modified boxplot of the paired differences.

Press 2nd STAT PLOT accessing the StatPlot menu, as shown in Figure 11.21. Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select L3 as the list, 2nd L3, as shown in Figure 11.22. Enter a frequency of 1 in Freq:. Press ZOOM 9: ZoomStat to view the modified boxplot as shown in Figure 11.23. Use the TRACE key to obtain the five-number summary of the data.

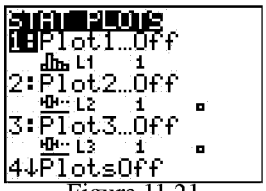


Figure 11.21

Figure 11.22

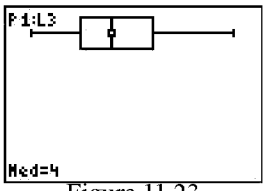


Figure 11.23

Figure 11.23 shows a modified boxplot of these differences. The sample size is under 30, so it is important to make sure there are no major outliers or extreme skewness. The sample mean and median are 5.36 and 4.0 hours, respectively. In data ranging from -20 to +40, these statistics are close enough to rule out extreme skewness. The boxplot of the differences in Figure 11.23 provides further evidence that the appropriate conditions are satisfied.

**Follow these steps** to construct a 90% confidence interval for the paired differences.

1. Select [STAT] [►] [►], accessing the TESTS menu. Select 8: TInterval, as shown in Figure 11.24, and press [ENTER]. Highlight Data and press [ENTER] to make it "stick". Use the down arrow key, [▼], to enter [2nd] [L3] in List:. Enter a frequency of 1 in Freq:. Type 0.90 in the confidence interval C-Level:. Use the down arrow key, highlighting Calculate and [ENTER] to execute the command, as shown in Figure 11.25. The results are shown in Figure 11.26.

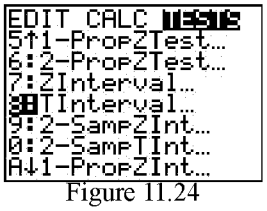


Figure 11.24

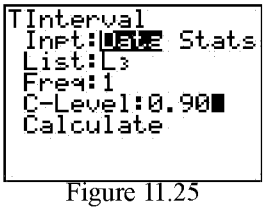


Figure 11.25

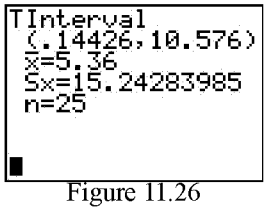


Figure 11.26

The TI output, shown in Figure 11.26, indicates the confidence interval (0.14426,10.576). The sample mean, $\overline{x}$, is 5.36 and the sample standard deviation, $s$, is 15.24, rounded to 2 decimal places. We are 90% confident that the average difference between computer usage and television viewing for students represented by this sample is covered by the interval from 0.14 to 10.58 hours per week. with more hours spent on computer usage than on television viewing.

### Interpreting a Confidence Interval for the Mean of Paired Differences

In addition to the usual interpretation of a confidence interval, it is often of interest to know whether the confidence interval for the mean of paired differences covers 0. If it does cover 0, then it is possible that the population mean is 0, indicating that the population means for the two measurements could be the same. If the interval does not cover 0, then we can be fairly certain that the population means for the two variables are different. How certain we can be depends on the confidence level used for the computation.

In Example 11.6, the 90% confidence interval for the mean difference in screen time for computer and TV was 0.14 hours to 10.58 hours per week. The interval does not cover 0, so we can be fairly certain that for the population represented by this sample students really do spend more time on computer usage than on watching television, on average.

## 11.4 Confidence Interval for the Difference in Two Means

### The General (Unpooled) Case

As is the case in confidence intervals for a single mean, the t-distribution is used to determine the multiplier in the confidence interval for the difference in means for independent samples. Because the multiplier is $t^*$ and two samples are used, the result is sometimes called a **two-sample t-interval**. The general format of a confidence interval for the difference in two means is

$$Difference\ in\ sample\ means \pm t^* \times Standard\ error$$

The *standard error* of the difference in sample means is

$$\text{s.e.}(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Unfortunately, there's a muddy mathematical story underneath the calculation of a confidence interval for the difference between two population means. On the surface, however, the story appears to be easy and can be summarized as follows. An approximate confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{x}_1 - \overline{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The multiplier $t^*$ is a t-value such that the area between $-t^*$and $+t^*$ in the appropriate $t$-distrlbution equals the desired confidence level. Appropriate degrees of freedom are difficult to specify. Computer software will provide the approximate df, but lacking that a conservative approximation is to use take the smaller of the two samele sizes and subtract 1.

**Two Situations for Which a**
**$t$ Confidence Interval for Two Means Is Valid**

An important condition required for the formulas in this section to be valid is that the samples must be *independent* of each other. In addition, one of two different situations must hold:

Situation 1: The populations of measurements are both bell-shaped, and random samples of any size are measured. In practice, for small samples the observed-data-should not show extreme skewness and there should not be any outliers.

Situation 2: Large random samples are measured. A somewhat arbitrary definition of a "large" sample is $n > 30$, but if there are extreme outliers, it is better to have an even larger sample.

Note: For a confidence Interval for the difference in two means, one of these situations must hold for both groups.

## The Equal Variance Assumption and the Pooled Standard Error

When estimating the difference between two population means, it may sometimes be reasonable to assume that the populations have equal standard deviations. The term variance describes the squared standard deviation, so the assumption of equal standard deviations for the two populations is the same as assuming the variances are equal. Using statistical notation, we can express the assumption of equal population variances as $\sigma_1^2 = \sigma_2^2 = \sigma^2$, where $\sigma^2$ represents the common value of the variance. With this assumption, information from both groups is combined in order to estimate the value of $\sigma^2$. The estimate of variance based on the combined or .pooled" data is called the pooled variance. The square root of the pooled variance is called the pooled standard deviation, and it is computed as

$$\text{Pooled standard deviation} = s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

Replacing the individual standard deviations $s_1$ and $s_2$ with the pooled version $s_p$ in the formula for standard error leads to the **pooled standard error for the difference between two means:**

$$\text{Pooled s.e.}(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

This may all seem quite complicated, but if the assumption of equal population variances is correct, this complication actually provides a cleaner mathematical solution for the determination of the multiplier $t^*$. In this case, the degrees of freedom are simply $df = n_1 + n_2 - 2$.

If we assume the population variances are the same, then the confidence interval for $\mu_1 - \mu_2$, the difference between the population means, is

$$(\overline{x}_1 - \overline{x}_2) \pm t^* \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$t^*$ is found using a t-distribution with $df = n_1 + n_2 - 2$, and $s_p$ is the pooled standard deviation.

### Example 11.14 - Pooled t-interval for Difference Beween Mean Female and Male Sleep Times

Students in an introductory statistics class filled out a survey on a variety of issues, including how much sleep they had the previous night. Let's use the data to estimate how much more or less male students sleep than female students, on average. The class included 83 females and 65 males. Let's assume these students are equivalent to a random sample of all students who take introductory statistics. How much difference is there between how long female and male students represented by this sample slept the previous night? To answer this question, we'll find a 95% confidence interval for $\mu_1 - \mu_2$ = difference in mean sleep hours for females

versus males. Here is the data.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3.50 | 4.00 | 5.00 | 8.50 | 7.00 | 6.50 | 5.00 | 9.00 | 9.00 |
| 6.50 | 5.50 | 7.00 | 9.50 | 7.00 | 6.50 | 5.00 | 6.00 | 7.00 |
| 7.00 | 5.50 | 7.50 | 9.00 | 7.00 | 6.00 | 8.50 | 7.00 | 11.00 |
| 8.00 | 4.00 | 7.00 | 4.50 | 4.00 | 7.00 | 9.50 | 9.00 | 9.00 |
| 8.00 | 8.50 | 6.50 | 5.00 | 5.25 | 7.50 | 8.00 | 6.50 | 6.00 |
| 10.00 | 7.00 | 8.00 | 8.50 | 6.00 | 7.00 | 4.00 | 7.00 | 10.00 |
| 8.00 | 7.00 | 6.00 | 10.00 | 4.50 | 6.00 | 3.00 | 5.00 | 4.50 |
| 7.00 | 9.00 | 8.00 | 8.00 | 8.00 | 6.50 | 6.00 | 10.00 | 8.00 |
| 9.00 | 7.00 | 8.00 | 4.50 | 8.00 | 9.50 | 8.00 | 8.00 | 8.00 |
| 5.00 | 6.00 | | | | | | | |

Table 11.4 Sleep Hours for Females

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9.0 | 6.0 | 6.0 | 9.0 | 8.0 | 4.5 | 8.0 | 8.0 | 5.0 |
| 5.0 | 4.0 | 8.5 | 5.0 | 8.5 | 9.0 | 6.0 | 6.5 | 5.0 |
| 6.0 | 7.0 | 8.0 | 7.0 | 6.5 | 6.0 | 6.5 | 9.0 | 4.0 |
| 3.5 | 4.5 | 5.0 | 8.0 | 7.5 | 8.0 | 2.0 | 9.0 | 7.5 |
| 12.0 | 8.0 | 5.0 | 5.0 | 6.5 | 6.5 | 7.0 | 7.0 | 7.0 |
| 7.0 | 6.0 | 8.0 | 5.0 | 6.0 | 6.0 | 7.5 | 5.5 | 7.0 |
| 9.0 | 7.0 | 6.0 | 6.0 | 5.5 | 7.0 | 5.5 | 6.0 | 4.0 |
| 7.0 | 5.0 | | | | | | | |

Table 11.5 Sleep Hours for Males

**Follow these steps** to construct two modified boxplots checking the conditions required for using the t confidence interval.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the data found in Table 11.4 and in Table 11.5.

      Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the data: 3.5, 4.0, 5.0 ... pressing [ENTER] after each entry, as shown in

Figure 11.27.
Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the data:9.0, 6.0, 6.0 ... pressing ENTER after each entry, as shown in Figure 11.27.

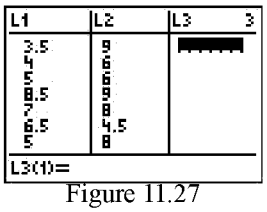


Figure 11.27

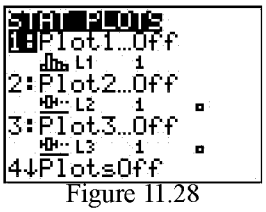


Figure 11.28

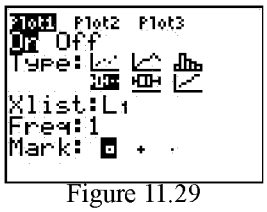


Figure 11.29

b. Plot the statistical data by creating two modified boxplots of the sleep times for females and males.

Press 2nd STAT PLOT accessing the StatPlot menu, as shown in Figure 11.28. Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select L1 as the list, 2nd L1, as shown in Figure 11.29. Enter a frequency of 1 in Freq:. Use the up arrow ▲, placing the cursor on Plot 2. Press ENTER, selecting Plot 2. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select L2 as the list, 2nd L2, as shown in Figure 11.30. Enter a frequency of 1 in Freq:. Press ZOOM 9: ZoomStat to view the two modified boxplots of the sleep times for females and males as shown in Figure 11.31. Use the TRACE key to obtain the five-number summaries of the data.

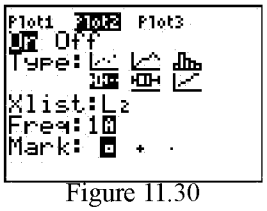


Figure 11.30

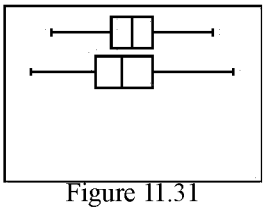


Figure 11.31

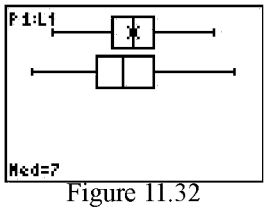


Figure 11.32

Figure 11.31 shows the two modified boxplots of the sleep times for females and males. The modified boxplots suggest that there are no major outliers or extreme skewness. The sample median for females is 7.0 and the sample median for males is 5.0 hours, respectively. In data ranging from 3 to 11 for females and from 2 to 12 for males, these statistics are close enough to rule out extreme skewness. The two modified boxplots of the sleep times for females and males in Figure 11.31 provides further evidence that the appropriate conditions are satisfied.

**Follow these steps** to construct a 95% confidence interval for $\mu_1 - \mu_2$ = difference in mean sleep hours for females versus males using a pooled standard deviation.

1. Select [STAT] [►] [►], accessing the TESTS menu. Select 0: 2-SampTInt..., as shown in Figure 11.33, and press [ENTER]. Highlight Data and press [ENTER] to make it "stick". Use the down arrow key, [▼], to enter [2nd] [L1] in List1:. Use the down arrow key, [▼], to enter [2nd] [L2] in List2:. Enter a frequency of 1 in Freq1:. Enter a frequency of 1 in Freq2:. Type 0.95 in the confidence interval C-Level:. Use the down arrow key to Pooled, highlighting **Yes**, and press [ENTER] to make it "stick". Use the down arrow key, Calculate and [ENTER] to execute the command, as shown in Figure 11.34. The results are shown in Figure 11.35.

Figure 11.33

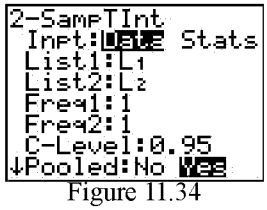


Figure 11.34

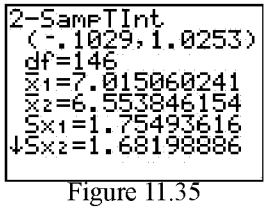


Figure 11.35

The TI output, shown in Figure 11.35, indicates the confidence interval is (-0.1029,1.0253).

Notice that the sample standard deviations are very similar (1.75 for females and 1.68 for males), so it may be reasonable to assume that the populatlon variances are similar as well. The sample mean given for females is $\overline{x}_1 = 7.02$ hours, the sample mean given for males is 6.55 hours with the difference between the sample means equal to 0.461.

The 95% confidence interval for the difference in mean hours of sleep for the populations of female and male students is -0.103 to 1.025 hours. Because the interval covers 0, we can't rule out the possibility that the population means are equal for men and women, although for this sample, women slept an average of about half an hour more than the men did. Use the down arrow, [▼], to view the pooled standard deviation given at the bottom of the output: $s_p = 1.72$.In the pooled procedure, the degrees of freedom are found as $df = n_1 + n_2 - 2$, which in this example equals 83 + 65 - 2 = 146.

**Pooled or Unpooled?**

In Example 11.14, the sample standard deviations for females and males had about the same vaiues, so it was reasonable to use the assumption that the population standard deviations were equal.

**Follow these steps** to construct a 95% confidence interval for $\mu_1 - \mu_2$ = difference in mean sleep hours for females versus males using a unpooled standard

deviation.

1. Select [STAT] [►] [►], accessing the TESTS menu. Select 0: 2-SampTInt..., as shown in Figure 11.36, and press [ENTER]. Highlight Data and press [ENTER] to make it "stick". Use the down arrow key, [▼], to enter [2nd] [L1] in List1:. Use the down arrow key, [▼], to enter [2nd] [L2] in List2:. Enter a frequency of 1 in Freq1:. Enter a frequency of 1 in Freq2:. Type 0.95 in the confidence interval C-Level:. Use the down arrow key to Pooled, highlighting **No**, and press [ENTER] to make it "stick". Use the down arrow key, Calculate and [ENTER] to execute the command, as shown in Figure 11.37. The results are shown in Figure 11.38.

Figure 11.36

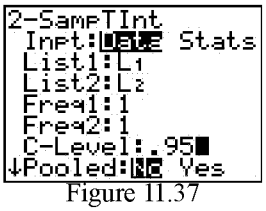


Figure 11.37

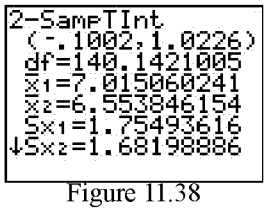


Figure 11.38

The TI output, shown in Figure 11.38, indicates the confidence interval is (-0.1002,1.0226). The confidence interval for the difference in means, however, would have been about the same even if the assumption of equal standard deviations had not been made. With the unpooled procedure, the 95% confidence interval for the difference in population means is -0.10 to 1.03 hours, quite close to the pooled version. One advantage for the pooled version is that finding the value of the degrees of freedom is simpler.

Sample standard deviations in two independent samples will almost never be identical in practice. So, how do we know when it might be reasonable to use the pooled version of a confidence interval for a difference between two population means? And, what is the risk of using the pooled procedure when, in truth, the population standard deviations differ? We will explore that question in detail when we consider hypothesis testing for the difference between two means in Chapter 13, but here we give some preliminary guidance.

If the larger value of the two sample standard deviations is from the group with the larger sample size, the pooled procedure will tend to give a wider confidence interval than the un pooled version and so would be a conservative estimate of the true difference. Similar to when we used the conservative margin of error in a confidence interval for one proportion, it is acceptable to use the more conservative pooled procedure. But, it is not good practice to knowingly create an interval that is wider than necessary. On the other hand, if the smaller of the two sample standard deviations is from the group with the larger sample size, the pooled version of the procedure may produce a misleading narrow interval. Generally,it's best to use the unpooled procedure unless the sample standard deviations are quite similar.

For this example, it is not reasonable to assume equal variances. Notice that do-

ing so increases the width of the confidence interval. Without the assumption the interval is from 0.23 to 1.46, but with the assumption it is 0.13 to 1.57. A comparison of the standard error of the difference for the two situations illustrates why the width increased. Using the pooled estimate increases the standard error of the mean for the sample with $n = 25$.

$$\text{Unpooled s.e.}(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(1.34)^2}{25} + \frac{(1.73)^2}{148}} = 0.30$$

$$\text{Pooled s.e.}(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{(1.68)^2}{25} + \frac{(1.68)^2}{148}} = 0.36$$

This example illustrates that it is conservative to use the pooled procedure if the two sample sizes are decidedly different and the larger sample standard deviation accompanies the larger sample size.

# Chapter 12
# Testing Hypotheses About Proportions

In Chapters 10 and 11 we learned how to find confidence intervals for population parameters in five different contexts involving proportions and means. In this chapter and the next one we will learn how to conduct tests of hypotheses for the same five population parameters.

Suppose that your state legislature is considering a proposal to lower the legal limit for the blood alcohol level that constitutes drunk driving. A legislator wants to determine if a majority of adults in her district favor this proposal. To gather information, she surveys 200 randomly selected individuals from her district and learns that 59% of these people favor the proposal.
Can the legislator conclude that a majority of all adults in her district favor this proposal? Because the result is based on a sample, there is the possibility that the observed majority might have occurred just by the "luck ofthe draw." If a majority ofthe whole population actually opposes the proposal, how likely is it that 59% of a. random sample would favor the proposal?
In this chapter, we will learn how to use the method of statistical hypothesis testing to analyze this type of issue. The hypothesis testing method uses data from a sample to judge whether or not a statement about a population may be true. For example, is it true that a majority of all adults in the legislator's district favor the proposal for the tougher blood alcohol standard?
Hypothesis testing and confidence interval estimation are related methods, and often both methods can be used to analyze the same situation. As we learned in Chapter 10, a confidence interval is a numerical answer to the question "What is the population value?" A hypothesis test is used to answer questions about particular values for a population parameter, or particular relationships in a population, based on information in the sample data.

We cover the basics for hypothesis testing for the five parameters of interest in Hypothesis Testing (HT) Module 0 in this chapter. In Modules 1 and 2 we provide details for the two situations involving population proportions, for which confidence interval methods were covered in Chapter 10. They include:

1. Hypotheses about one population proportion. For example, full moons occur every on average every 29.53 days, and we might classify babies as to whether they were born during the 24 hour period surrounding a full moon or not. The hypothesis of interest is that the proportion of babies born during a full moon is 1/29.53, as it would be if births were uniformly distributed across days.
2. Hypotheses about the difference between two population proportions. For example, we might question whether or not the proportion favoring the death

penalty is the same for teenagers as it is for adults.

After reading this chapter you should be able to:

1. Test hypotheses about a proportion using summarized data where the sample proportion is given.
2. Find an exact binomial p-value when testing one proportion.
3. Find power for specified true parameters and sample sizes.
4. Conduct a two-sample $z$-test for the difference in proportions using summarized data.

**Keystrokes Introduced**

1. [STAT] [►] [►], accessing the TESTS menu and selecting 5: 1-PropZTest. **1-PropZTest** computes a test for an unknown proportion of successes (prop). It takes as input the count of successes in the sample $x$ and the count of observations in the sample $n$. **1-PropZTest** test the null hypothesis $H_0$: $prop = p_0$ against one of three possible alternatives: $H_a$: $prop \neq p_0$, $H_a$: $prop < p_0$, or $H_a$: $prop > p_0$.
2. [2nd] [DISTR], accessing the DISTR menu and selecting A: binomcdf(. The **binomcdf**(*numtrials*,*p* [,*x*]) function computes a cumulative probability at *x* for the discrete binomial distribution with the specified *numtrials* and probability $p$ of success on each trial.
3. [2nd] [DISTR], selecting 2: normalcdf(. The arguments are
   2: normalcdf(*lowerbound*,*upperbound* [$\mu$,$\sigma$]).
   The command computes the cumulative normal distribution probability between *lowerbound* and *upperbound* for the specified $\mu$ and $\sigma$. If $\mu$ and $\sigma$ are not specified, the default is the standard normal distribution.
4. [STAT] [►] [►], accessing the TESTS menu and selecting 6: 2-PropZTest. **2-PropZTest** computes a two-proportion $z$ test. The arguments in the two-proportion $z$ test are 5: 1-PropZTest ($x_1$,$n_1$,$x_2$,$n_2$[,*alternative* $<$, $\neq$, $>$„*drawflag (draws results, calculates results)]*. **2-PropZTest** test the null hypothesis $H_0$: $prop = p_0$ against one of three possible alternatives: $H_a$: $prop \neq p_0$, $H_a$: $prop < p_0$, or $H_a$: $prop > p_0$.

## 12.1 Testing Hypotheses About a Population Proportion

In this section, we illustrate hypothesis testing in the context of testing a specified value for a population proportion. We encountered this situation in Chapter 8 when we learned about binomial random variables. In that case we focused on $X$ = number of successes, whereas in this section we focus on the sample proportion, $\frac{X}{n}$. However, the scenarios in which the methods apply are equivalent, and the population parameter $p$ in this section is equivalent to the probability of success $p$

in a binomial setting.

### The Z- Test for a Proportion

When we have a sufficiently large random sample, we can use a **z-test** to examine hypotheses about a population proportion. The test is called a "$z$-test" because the **test statistic** is a standardized score ($z$-score) for measuring the difference between the sample proportion and the null hypothesis value of the population proportion. The logic of the test is as follows:

1. Determine the sampling distribution of possible sample proportions when the true population proportion is $p_0$ (called the null value), the value specified in $H_0$.
2. Using properties of this sampling distribution, calculate a standardized score ($z$-score) for the observed sample proportion $\widehat{p}$.
3. If the standardized score has a large magnitude, conclude that the sample proportion $\widehat{p}$ would be unlikely if the null value $p_0$ is true, and reject the null hypothesis.

The principal idea is that we determine how likely it would have been for the sample proportion to have occurred if the null hypothesis is true.

Conditions for Conducting the z- Test

Whenever we conduct a hypothesis test, we must make sure the data meet certain conditions. These conditions are the assumptions that were made when the theory was derived for the test statistic that we are using to test our hypotheses. For testing proportions with the procedure described in this section, there are two conditions that should be true for the sample:

1. The sample should be a random sample from the population.
2. The quantities $np_0$ and $n(l - p_0)$ should both be at least 10.

**Example 12.6 - The Importance of Order**

This example was first introduced in Chapter 2. In a student survey done in a statistics class, 92 college students were given a form read "Randomly choose one of the letters S or Q." Another 98 students were given a form with the order ofthe letters reversed, to read "Randomly choose one of the letters Q or S." The purpose was to determine if the order of listing the letters might influence the choice of letters. The possible influence of the order of listing items is a concern in elections. Many election analysts feel a candidate gains an advantage if he or she is the first candidate listed on the ballot. The data is contained in the *PennState1* worksheet. Among 92 students who saw the order" S or Q," 61 picked S, the first choice. Among the 98 students who saw the order "Q or S," 53 picked Q, the first choice. In all, 114 of 190 students picked the first choice of letter. Expressed as a

proportion, this is $\widehat{p} = \frac{114}{190} = 0.60$.

Use the letter p to represent the proportion of the population that would pick the first letter. the null hypothesis is a statement of "nothing happening." If there is no general preference for either the first or second letter, $p = 0.5$ (because there are two choices). So the "null value" for $p$ is $p_0 = 0.5$.
The alternative hypothesis ususally states the researcher's belief or speculation. The purpose of the acitivity was to see if there is a general preference for picking the first choice, and a preference for the first letter would mean that $p$ is greater than 0.5. The null and alternative hypotheses can be summarized as

$$H_0 : p = 0.5$$
$$H_a : p > 0.5$$

**Verify necessary data conditions**, and if met, summarize the data into an appropriate test statistic. Before computing the test statistic, we verify that our sample meets the necessary conditions for using the z-statistic. With $n = 190$ and $p_0 = 0.5$, both $np_0$ and $n(l - p_0)$ equal 95, a quantity larger than 10, so the sample size condition is met. The sample, however, is not really a random sample-it's a convenience sample of students who were enrolled in this class. It does not seem that this will bias the results for this question, so we will behave as though the sample was a random sample.

**Follow these steps** to conduct a hypothesis test about a proportion using summarized data where the sample proportion is given.

1. Conduct the one-sample $z$-test about a proportion using summarized data where the sample proportion is given.
   Select STAT ▶ ▶, accessing the TESTS menu. Select 5: 1-PropZTest, as shown in Figure 12.1, and press ENTER. Type the value of the null hypothesis in $p_0$: 0.5. Use the down arrow key, ▼, to enter 114 in $x$:. Use the down arrow key, ▼, to enter 190 in $n$:. Use the down arrow key, ▼, to highlight the alternative hypothesis $> p_0$ and press ENTER to make it "stick". Use the down arrow key, highlighting Calculate and ENTER to execute the command, as shown in Figure 12.2. The results are shown in Figure 12.3.

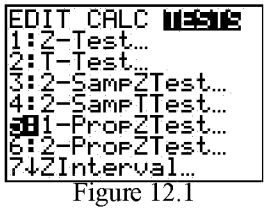


Figure 12.1

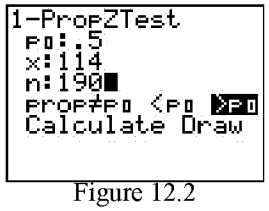


Figure 12.2

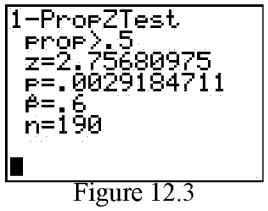


Figure 12.3

   The TI output, as shown in Figure 12.3, indicates the value of the test statistic, $z = 2.76$, the $p$-value, $p = 0.003$, and the sample proportion, $\widehat{p} = 0.6$. The statistical conclusion was to reject the null hypothesis that p = .50. In this situation, the conclusion is that there is statistically significant evidence that the first letter presented is preferred. This conclusion is a generalization that applies to

the population represented by this sample. In other words, we conclude that for the entire population of individuals represented by these students, there would be a preference for choosing the first letter presented.

## Exact p-values In Tests for Population Proportions

The method for finding a p-value shown in Table 12.1 of the text relies on the fact that the test statistic $z$ is approximately normal. However, in the special case in this Section, of testing one population proportion, there is a method available for finding an exact p-value. Recall that the $p$-value is the *probability of a result as extreme (or more extreme) as the observed test statistic, in the direction of the alternative hypothesis, if the null hypothesis is true*.
Studies designed to test a population proportion are equivalent to binomial experiments. Although we have focused on the sample proportion of successes $p = \frac{X}{n}$, we could equivalently focus on the number of successes $X$. Let's look at an example.

**Example 12.8 Making Sure Students Aren't Just Guessing**

Suppose a teacher gives a true/false test with 10 questions. How can she reject the hypothesis that students are just guessing? The parameter of interest is $p$ = probability of getting each question correct, and under the null hypothesis that students are just guessing, this is a binomial experiment with $n = 10$ and $p = 0.5$.The test of interest is:

$$H_0 : p = 0.5$$
$$H_a : p > 0.5$$

How many questions would a student need to get right before the teacher rejects the null hypothesis using the standard $\alpha = .05$? Suppose a students gets $X = 7$ questions correct. Then we can find the $p$-value based on our knowledge of binomial probabilities from Chapter 8. The $p$-value = probability of observing 7 or more correct answers if the student is guessing = $P(X = 7$ or $8$ or $9$ or $10)$.

**Follow these steps** to find the $p$-value = probability of observing 7 or more correct answers.

1. The **binomcdf**(*numtrials,p* [,*x* ]) function computes a cumulative probability at $x$ for the discrete binomial distribution with the specified *numtrials* and probability $p$ of success on each trial. You will subtract the probability of observing 6 or fewer correct answers from 1 to find the $p$-value = probability of observing 7 or more correct answers. Type 1 [−] [2nd] [DISTR], selecting A: binomcdf(, as shown in Figure 12.4, and press [ENTER]. Type the number of trials, 10, a comma, [,], the probability of success, 0.5, and 6, the probability of observing 6 or fewer correct answers, followed by [)], as shown in Figure 12.5. Press

ENTER to execute the command. The results are shown in Figure 12.6.

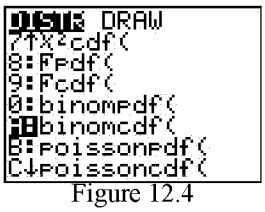


Figure 12.4

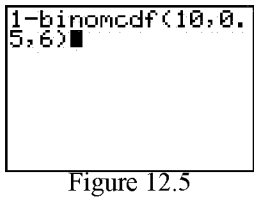


Figure 12.5

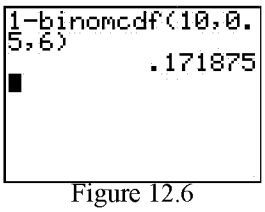


Figure 12.6

The TI output, as shown in Figure 12.6, indicates if a student scored 7 correct on the exam, the p-value is .17187, and the null hypothesis should not be rejected.

Suppose the student scored 8 answers correct. The $p$-value = probability of observing 8 or more correct answers if the student is guessing = $P(X = 8$ or 9 or 10).

**Follow these steps** to find the $p$-value = probability of observing 8 or more correct answers.

1. The **binomcdf(***numtrials*,*p* [,*x*]) function computes a cumulative probability at $x$ for the discrete binomial distribution with the specified *numtrials* and probability $p$ of success on each trial. You will subtract the probability of observing 7 or fewer correct answers from 1 to find the $p$-value = probability of observing 8 or more correct answers. Type 1 [−] [2nd] [DISTR], selecting A: binomcdf(, as shown in Figure 12.7, and press [ENTER]. Type the number of trials, 10, a comma, [,], the probability of success, 0.5, and 7, the probability of observing 7 or fewer correct answers, followed by [)], as shown in Figure 12.8. Press [ENTER] to execute the command. The results are shown in Figure 12.9.

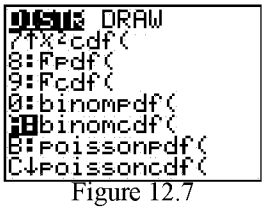


Figure 12.7

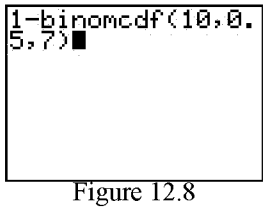


Figure 12.8

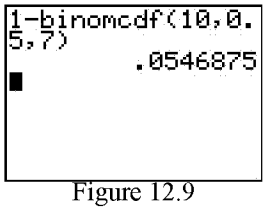


Figure 12.9

The TI output, as shown in Figure 12.9, indicates if a student scored 8 correct on the exam, the p-value is .05469, and the null hypothesis should not be rejected. In this case the p-value is .05469, still not quite enough to reject the null hypothesis using .05 as the level of significance. A student would need to score 9 or 10 correct to provide convincing evidence that he or she was not just guessing. Fortunately, most true/false tests are not graded using this criterion for a passing grade!

## 12.2 Power and Sample Size For Testing One Proportion

Remember that the power of a statistical test is the probability that the null hypoth-

esis will be rejected, given a specific true value for the population parameter. For instance, we learned that the null hypothesis will be rejected in a one-tailed test (with "greater than" as the alternative and $\alpha = .05$) if the test statistic is 1.645 or larger. So, the power of a test is the probability that the test statistic will be 1.645 or higher, given a specific true value for the population parameter.
For a specified true value and level of significance, we can calculate the power for a given sample size, or we can compute the sample size required to achieve specified power. Researchers often investigate this while they are planning a study, and plan to take a sample large enough to achieve power of .80 or better, for a reasonable guess of what the true population parameter value is.

**Example 12.3 Planning a Student Survey**

A university is considering a plan to offer regular classes year-round by making the summer session a regular term but will only do so if there is sufficient student interest. A university administrator is planning a student survey to determine if a majority of all students at the school would attend summer session under that structure. The plan will only be implemented if a majority would attend so the status quo is to assume that the proportion who would attend is not a majority. With $p$ representing the proportion of all students who would attend in the summer, a hypothesis testing structure is

$H_0 : p \leq 0.5$ (the proporiton who would attend is not a majority)
$H_a : p > 0.5$ (a majority would attend)

Notice that the alternative hypothesis includes a broad range of possibilities and does not specify an exact value. If the true population proportion who would attend is any value greater than .50, the correct decision is to pick the alternative hypothesis.

Suppose the true population proportion of 0.60 is correct. The power is the probability that the sample evidence leads us to conclude that a majority of the student population would attend a regular summer term.

**Follow these steps** to find the power for specified true parameters and sample sizes.

1. Find the appropriate $z$-multiplier for a 0.05 level of significance using a "one-tailed" alternative.

   In order to find the appropriate $z$-multiplier for a 0.05 level of significance using a "one-tailed" alternative, we use the invNorm function requiring the area to the left of the $z$-multiplier. Therefore for a 0.05 level of significance using a "one-tailed" alternative, you will have 5% (0.05) in the upper tail of the standard normal curve, as shown in Figure 12.10. The area to the left of the positive $z$-multiplier will then be 0.95. Take the following steps:
   On the homescreen, press [2nd] [DISTR], located on the fourth row, column four, above [VARS]. Select 3: invNorm( .Type 0.95 [)] [ENTER]. The results

are shown in Figure 12.11, indicating the appropriate $z$-multiplier for a 0.05 level of significance using a "one-tailed" alternative is 1.645, rounded to 3 decimal places.

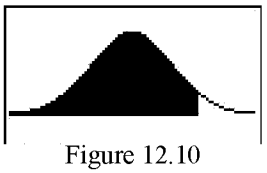


Figure 12.10

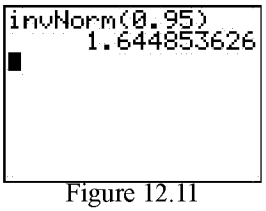



Figure 12.11

a. Find the power for power for three different "true" proportions and for three different sample sizes.

(i) Find the power for power if the "true" proportion is 0.52 and the sample sizes are 50, 100, and 400.

(A) Find the power given that the "true" proportion is 0.52 and the sample size is 50.

What we will do is use the normalcdf( function where the arguments are *lowerbound,upperbound*. The *lowerbound* and *upperbound* will be indicated in terms of a $z$-statistic. For the *lowerbound*, the $z$-statistic is found by using

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(0.5 + 1.645 * \sqrt{\frac{0.5\times 0.5}{50}} - 0.52}{\sqrt{\frac{0.5\times 0.5}{50}}}$$

where $\widehat{p}$ is found by taking the value contained in the null hypothesis and adding $z^* \times null\ standard\ error$, with the sample size being $n = 50$. The *upperbound* will be a large value of $z$, $1E99$. Press 2nd DISTR, selecting 2: normalcdf( and press ENTER. Type another parenthesis, ( and 0.5+1.645 × 2nd $\sqrt{}$ 0.5 × 0.5 ÷ 50 ) − 0.52 ) ÷ 2nd $\sqrt{}$ 0.5 × 0.5 ÷ 50 ) , 1 2nd EE 99, as shown in Figure 12.12. Press ENTER to obtain the power.

(B) Find the power given that the "true" proportion is 0.52 and the sample size is 100.

Press 2nd ENTER to retrieve the previous keystrokes. Use the up arrow, ▲, placing the cursor on the 5 of each of the sample sizes of 50. Change the 5 to a 1 and press 2nd INS and type 0, to change the sample size of 50 to a sample size of 100, as shown in Figure 12.13

(C) Find the power given that the "true" proportion is 0.52 and the sample size is 400.

Press 2nd ENTER to retrieve the previous keystrokes. Use the up arrow, ▲, placing the cursor on the 1 of each of the sample sizes of 100. Change the 1 to a 4 to change the sample size of 100 to a sample size of 400, as shown in Figure 12.14.

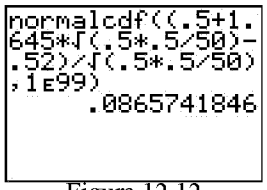


Figure 12.12

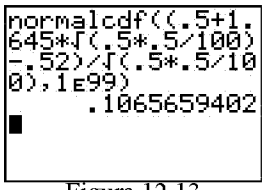


Figure 12.13

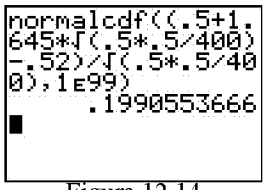


Figure 12.14

Let us examine the TI output, as shown in Figures 12.12, 12.13 and 12.14, for all three calculations. If in truth 52% of students think they would attend a regular summer term and $n = 50$ students are surveyed, the probability that the sample proportion would be large enough to conclude that $p > 0.50$ is 0.09 (power = 0.09). That means the probability is 0.36 that a type 2 error would be made, and the null hypothesis would not be rejected. If the sample size is $n = 100$, the power is 0.11; if the sample size is $n = 400$, the power is 0.20.

(ii) Find the power for power if the "true" proportion is 0.60 and the sample sizes are 50, 100, and 400.

(A) Find the power given that the "true" proportion is 0.60 and the sample size is 50.

What we will do is use the normalcdf( function where the arguments are *lowerbound,upperbound*. The *lowerbound* and *upperbound* will be indicated in terms of a $z$-statistic. For the *lowerbound*, the $z$-statistic is found by using

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(0.5 + 1.645 * \sqrt{\frac{0.5\times0.5}{50}} - 0.60}{\sqrt{\frac{0.5\times0.5}{50}}}$$

where $\widehat{p}$ is found by taking the value contained in the null hypothesis and adding $z^* \times null\ standard\ error$, with the sample size being $n = 50$. The *upperbound* will be a large value of $z$, $1E99$.
Press 2nd DISTR, selecting 2: normalcdf( and press ENTER.
Type another parenthesis, ( and 0.5+1.645 × 2nd √ 0.5 × 0.5 ÷ 50 ) − 0.60 ) ÷ 2nd √ 0.5 × 0.5 ÷ 50

) , 1 2nd EE 99, as shown in Figure 12.15. Press ENTER to obtain the power.

(B) Find the power given that the "true" proportion is 0.60 and the sample size is 100.

Press 2nd ENTER to retrieve the previous keystrokes. Use the up arrow, ▲, placing the cursor on the 5 of each of the sample sizes of 50. Change the 5 to a 1 and press 2nd INS and type 0, to change the sample size of 50 to a sample size of 100, as shown in Figure 12.16

(C) Find the power given that the "true" proportion is 0.60 and the sample size is 400.

Press 2nd ENTER to retrieve the previous keystrokes. Use the up arrow, ▲, placing the cursor on the 1 of each of the sample sizes of 100. Change the 1 to a 4 to change the sample size of 100 to a sample size of 400, as shown in Figure 12.17.

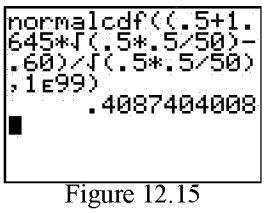


Figure 12.15

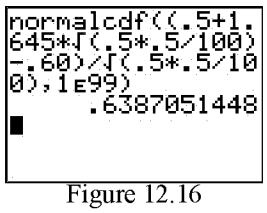


Figure 12.16

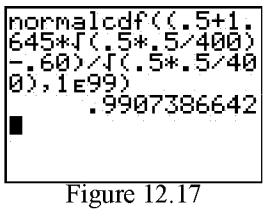


Figure 12.17

Let us examine the TI output, as shown in Figures 12.15, 12.16, and 12.17, for all three calculations. If in truth 60% of students think they would attend a regular summer term and $n = 50$ students are surveyed, the probability that the sample proportion would be large enough to conclude that $p > 0.50$ is 0.41 (power = 0.41). That means the probability is 0.59 that a type 2 error would be made, and the null hypothesis would not be rejected. If the sample size is $n = 100$, the power is 0.64; if the sample size is $n = 400$, the power is 0.99.

(iii) Find the power for power if the "true" proportion is 0.65 and the sample sizes are 50, 100, and 400.

(A) Find the power given that the "true" proportion is 0.65 and the sample size is 50.

What we will do is use the normalcdf( function where the arguments are *lowerbound,upperbound*. The *lowerbound* and *upperbound* will be indicated in terms of a $z$-statistic. For the

*lowerbound*, the $z$-statistic is found by using

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(0.5 + 1.645 * \sqrt{\frac{0.5\times0.5}{50}} - 0.65}{\sqrt{\frac{0.5\times0.5}{50}}}$$

where $\widehat{p}$ is found by taking the value contained in the null hypothesis and adding $z^* \times null\ standard\ error$, with the sample size being $n = 50$. The *upperbound* will be a large value of $z$, $1E99$. Press [2nd] [DISTR], selecting 2: normalcdf( and press [ENTER]. Type another parenthesis, [(] and 0.5+1.645 [×] [2nd] [√] 0.5 [×] 0.5 [÷] 50 [)] [−] 0.65 [)] [÷] [2nd] [√] 0.5 [×] 0.5 [÷] 50 [)] [,] 1 [2nd] [EE] 99, as shown in Figure 12.18. Press [ENTER] to obtain the power.

(B) Find the power given that the "true" proportion is 0.65 and the sample size is 100.

Press [2nd] [ENTER] to retrieve the previous keystrokes. Use the up arrow, [▲], placing the cursor on the 5 of each of the sample sizes of 50. Change the 5 to a 1 and press [2nd] [INS] and type 0, to change the sample size of 50 to a sample size of 100, as shown in Figure 12.19

(C) Find the power given that the "true" proportion is 0.60 and the sample size is 400.

Press [2nd] [ENTER] to retrieve the previous keystrokes. Use the up arrow, [▲], placing the cursor on the 1 of each of the sample sizes of 100. Change the 1 to a 4 to change the sample size of 100 to a sample size of 400, as shown in Figure 12.20.

```
normalcdf((.5+1.
645*√(.5*.5/50)-
.65)/√(.5*.5/50)
,1E99)
      .6830769065
```

Figure 12.18

```
normalcdf((.5+1.
645*√(.5*.5/100)
-.65)/√(.5*.5/10
0),1E99)
      .912291162
```

Figure 12.19

```
normalcdf((.5+1.
645*√(.5*.5/400)
-.65)/√(.5*.5/40
0),1E99)
      .9999933415
```

Figure 12.20

Let us examine the TI output, as shown in Figures 12.18, 12.19, and 12.20, for all three calculations. If in truth 65% of students think they would attend a regular summer term and $n = 50$ students are surveyed, the probability that the sample proportion would be large enough to conclude that $p > 0.50$ is 0.69 (power = 0.69). That means the probability is 0.319 that a type 2 error would be made, and the null hypothesis would not be rejected. If the sample size is $n = 100$, the power is 0.91; if the sample size is $n = 400$, the power is almost 1.

This example illustrates that power is a function of multiple factors. What patterns can we see? You should be able to detect two important relationships that apply to power for all hypothesis tests:

1. *The **power** increases when the **sample size** is increased.* We can see this by examining how power increases as the sample size increases. This makes sense because when the sample size is increased, the standard error is decreased, leading to larger values of the test statistic. Also, the sample statistic is .a more accurate estimate of the population value, making it easier to detect a difference between the true population value and the null value.
2. *The power increases when the difference between the true population value and the null hypothesis value increases.* We can see this when the sample size is held constant and the distance between the true population value and the null hypothesis value increases.This makes sense because the probability of detecting.a large difference is higher than the probability of detecting a small difference. However,remember that the truth about the population is not something the researcher can control or change.

Researchers should evaluate power before they collect data that will be used to do hypothesis tests to make sure they have sufficient power to make the study worthwhile. In this example,we see that a sample of 50 students would have low power even if the true proportion who would attend in the summer were as high as .60 or even .65, a definite majority. ITthis hypothesis test is important to the administrator for making a decision, he or she should sample more than 50 students. We also see that if the true population proportion is .52, only a slight majority, there is little chance that the administrator will be able to decide in favor of the alternative, even if 400 students are surveyed.

## 12.3 Testing Two Population Proportions

Researchers often wish to examine the difference between two populations with regard to the proportions falling into a particular category of a response variable. It is almost always of interest to test whether the two population proportions are equal, and that is the only case we will consider. The two populations may be represented by two categories of a categorical variable, such as when we want to compare the proportions of Republicans and Democrats who support a certain political issue, Or, the two populations may be hypothetical, represented by different treatments in an experiment. For instance, we may want to compare the proportions of smokers who would quit smoking if wearing a nicotine patch versus ifwearing a placebo patch.

When two proportions are equal, the standard errors of the sample statistics (the sample proportions) are equal as well.This fact will be used to estimate the null standard error for the test statistic. The null standard error formula used for the test differs from the standard error formula used in Chapter 11 to find confidence intervals for the difference in two proportions. Also, it is appropriate to use the phrase "null standard error" rather than simply "standard error" because the cal-

culation is connected to assuming the null hypothesis to be true.

**Example 12.10 - The Prevention of Ear Infections**

Based on its biochemical properties, Finnish researchers hypothesized that regular use of the sweetener xylitol might be useful for preventing ear infections in preschool children and carried out a study to test this hypopthesis (Uhari, 1998). In a randomized experiment, $n_1 = 165$ children took five daily doses of placebo syrup, and 68 of these children got an ear infection during the three months of the study. Another $n_2 = 159$ children took five daily doses of xylitol, and 46 of these children got an ear infection during the study. The sample proportions getting an ear infection are $\widehat{p}_1 = \frac{68}{165} = 0.412$ for the placebo group and $\widehat{p}_2 = \frac{46}{159} = 0.289$ for the xylitol group, and the difference between these two proportions is $0.123$ (12.3%). Is this observed difference in proportions large enough to conclude in general that using xylitol reduces risk of ear infection?
The researchers hoped to show that xylitol reduces ear infections. The parameter $p_1$ is the proportion who would get an ear infection in the population of children similar to those in the study if taking the placebo. The parameter $p_2$ is the proportion who would get an ear infection in the population of children similar to those in the study if taking xylitol. Therefore, the null and alternative hypotheses are

$$H_0 : p_1 - p_2 = 0 \text{ or } H_0 : p_1 = p_2$$
$$H_a : p_1 - p_2 > 0 \text{ or } H_a : p_1 > p_2$$

A one-sided alternative is used because the researchers want to show that the proportion getting an ear infection is significantly lower in the xylitol group than in the placebo group.

**Follow these steps** to verify the necessary data conditions required for using the two-sample $z$-test.

1. The example indicates that independent samples are available from the two populations.
2. The number with the response of interest and the number with the response of interest is at least 5 in each sample, and preferably at least 10.

   There are at least 10 children in each sample who did and did not get ear infections, so the conditions are met.

**Follow these steps** to conduct a two-sample $z$-test for the difference in proportions.

1. Conduct the two-sample $z$-test for the difference in proportions.
   Select [STAT] [►] [►], accessing the TESTS menu. Select 6: 2-PropZTest, as shown in Figure 12.21, and press [ENTER]. Use the down arrow key, [▼], to enter 68 in $x_1$:. Use the down arrow key, [▼], to enter 165 in $n_1$:. Use the down arrow key, [▼], to enter 46 in $x_2$:. Use the down arrow key, [▼], to enter 159 in

$n_2$:. Use the down arrow key, ▼, to highlight the alternative hypothesis $> p_2$ and press ENTER to make it "stick". Use the down arrow key, highlighting Calculate and ENTER to execute the command, as shown in Figure 12.22. The results are shown in Figure 12.23.

Figure 12.21

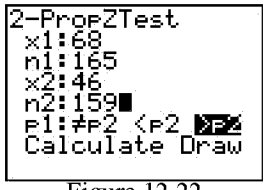


Figure 12.22

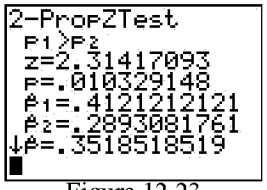


Figure 12.23

The TI output, shown in Figure 12.23, indicates the $Ha : p_1 > p_2$; the test statistic, $z = 2.31$; the $p$-value of the test statistic is $0.010$. The sample proportion for the placebo group is $\widehat{p}_1 = 0.412$; the sample proportion for the xylitol group is $\widehat{p}_2 = 0.289$. Using the down arrow key, ▼, reveals the the combined proportion is $\widehat{p}_1 = 0.35$. The sample size for the placebo group is $n_1 = 165$; the sample size for the xylitol is $n_2 = 159$. Using a = 0.05 as the level of significance criterion, the results are statistically significant because $0.01$, the p-value of the test, is less than $0.05$. In other words, we can reject the null hypothesis.
Based on this experiment, we can concluded that taking xylitol would reduce the proportion of ear infections in the populatioon of similar preschool children in comparison to taking a placebo.

# Chapter 13
# Testing Hypotheses About Means

In this chapter we continue our discussion of hypothesis testing, initiated in Chapter 12.

## 13.1 Introduction to Hypotheses Tests For Means

In Chapter 12, we leamed the logic and basic steps of hypothesis testing and applied them to z-tests for one proportion and the difference in two proportions. In this chapter, we'll look at the details of hypothesis tests for the same three situations for which we developed confidence intervals in Chapter 11, all of which involve population means:

1. Hypotheses about one population mean, $\mu$.For example, we might wish to determine whether or not the mean body temperature for humans is 98.6 degrees Fahrenheit.
2. Hypotheses about the population mean difference for paired data, $\mu_d$. For example, we might ask if the mean difference in height between college men and their fathers is 0.The data for each pair is "son's height - father's height" and we are interested in the mean of those differences for the population of father-son pairs in which the "son" is a college student.
3. Hypotheses about the difference between the means of two populations, $\mu_1 - \mu_2$. For example, we might ask whether the mean time of television watching per week is the same for men and women.

   Tests to answer these kinds of questions are sometimes called significance tests.

The terms **hypothesis testing** and **significance testing** are synonymous. The term *significance testing* arises because the conclusion about whether to reject a null hypothesis is based on whether a sample statistic is "significantly" far from what would be expected if the null hypothesis were true in the population. Equivalently, we declare **statistical significance** and reject the null hypothesis if there is a relatively small probability that an observed difference or relationship in the sample would have occurred if the null hypothesis holds in the population. One major principle for all significance tests is that we declare "statistical significance" when the p-value of the test is "small."
The significance tests for the three situations that we examine in this chapter all have the same general format because the test statistic is a standardized score that measures the difference between an observed statistic and the null value of a parameter. This was also the case in Chapter 12 where we used a $z$-statistic to test

hypotheses about a population proportion. The basic format that we followed in Chapter 12 will also be followed in this chapter.

After reading this chapter you should be able to:

1. Check the conditions before conducting a one-sample $t$-test.
   a. Conduct a one-sample $t$-test.
   b. Find the $p$-value for a one-sample $t$-test.
2. Check the conditions before conducting a paired $t$-test.
   a. Conduct a paired $t$-test.
3. Check the conditions before conducting a two-sample $t$-test.
   a. Conduct a two-sample $t$-test-unpooled.
   b. Conduct a two-sample $t$-test-pooled.

**Keystrokes Introduced**

1. [STAT] [►] [►], accessing the TESTS menu and selecting 2: T-Test. **T-Test** ... performs a $t$ test with frequency *freqlist*, *alternative* $<, \neq, >$, *drawflag* (draws results, calculates results). The arguments are 2: T-Test $\mu_0$ [,*listname, freqlist,alternative, drawflag]* (Data list input).
2. [2nd] [DISTR], selecting 5: tcdf( , the Student's-$t$ cumulative distribution probability function. The arguments are tcdf(*lowerbound*, *upperbound*, *df*).This function computs the Student's-$t$ distribution probability between *lowerbound* and *upperbound* for the specified degrees of freedom *df*.
3. [STAT] [►] [►], accessing the TESTS menu and selecting 4: 2-SampTTest. **2-SampTTest...** computes a two-sample $t$ test. *Pooled* = Yes polls variances; *Pooled* = No does not pool variances. The arguments are 2-SampTTest... (*listname1,listname2, freqlist1,freqlist2,, alternative* $<, \neq, >$*,pooled drawflag* (draws results, calculates results)*]*.

## 13.2 Testing Hypotheses About One Mean

For questions about **the mean** of a quantitative variable **for one population**, the null hypothesis typically has the form

$$H_0 : \mu = \mu_0 \qquad \text{(mean is a specified value)}$$

The alternative may either be one-sided ($H_a : \mu < \mu_0$ or $H_a : \mu > \mu_0$) or two-sided ($H_a : \mu \neq \mu_0$), depending on the research question of interest. The usual procedure for testing these hypotheses is called a **one-sample t-test** because the $t$-distribution is used to determine the $p$-value. Let's look at an example of a situation where a one-sample $t$-test would be used.

**Example 13.1 Normal Human Body Temperature**
What is normal body temperature? Apaper published in the *Journal of the American Medical Association* presented evidence that normal body temperature may be less than 98.6 degrees Fahrenheit, the long-held standard (Mackowiak et aI.,1992).The value 98.6 degrees seems to have come from determining the mean in degrees Celsius, rounding up to the nearest whole degree (37 degrees), and then converting that number to Fahrenheit using 32 + (1.8)(37) = 98.6. Rounding up may have produced a result higher than the actual average, which may therefore be lower than 98.6degrees. To test this, the null hypothesis is $\mu = 98.6$, and we are only interested in rejecting in favor of lower values, so the alternative is $\mu < 98.6$. We can write these hypotheses as

$$H_0 : \mu = 98.6$$
$$H_a : \mu < 98.6$$

Suppose that a random sample of $n = 18$ normal body temperatures is

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 98.2 | 97.8 | 99.0 | 98.6 | 98.2 | 97.8 | 98.4 | 99.7 | 98.2 |
| 97.4 | 97.6 | 98.4 | 98.0 | 99.2 | 98.6 | 97.1 | 97.2 | 98.5 |

Table 13.1

**Follow these steps** to construct a modified boxplot checking the conditions required before conducting a one-sample $t$-test..

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Turn off all StatPlots.

   Press [2nd] [STAT PLOT] accessing the StatPlot menu. Select 4: PlotsOff, pressing [ENTER] and [ENTER] again to turn all plots off.
   c. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Enter the data found in Table 13.1.

   Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the data: 98.2, 97.8, 99.0, ... pressing [ENTER] after each entry, as shown

in Figure 13.1.

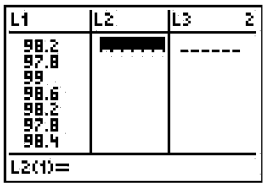



Figure 13.1

b. Plot the statistical data by creating a modified boxplot of the randomly selected sample of $n = 18$ normal body temperatures.

Press 2nd STAT PLOT accessing the StatPlot menu, as shown in Figure 13.2. Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select L1 as the list, 2nd L1, as shown in Figure 13.3. Enter a frequency of 1 in Freq:. Press ZOOM 9: ZoomStat to view the modified ooxplot as shown in Figure 13.4.

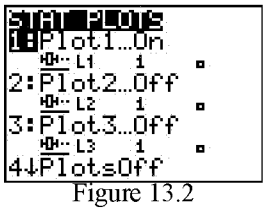



Figure 13.2

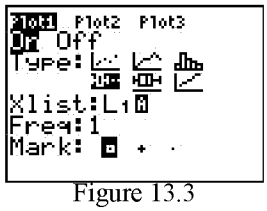



Figure 13.3

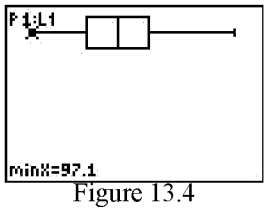



Figure 13.4

Figure 13.4 shows a modified boxplot of these normal body temperatures. The boxplot in Figure 13.4 illustrates that there are no outliers and the shape does not appear to be notably skewed. Although not shown on the boxplot, the sample mean = $\overline{x} = 98.217$ is almost identical to the median of 98.2. This comparison provides additional evidence that skewness and outliers are not a problem, so the one-sample t-test can be used.

Press 2nd STAT PLOT accessing the StatPlot menu. Select 4: PlotsOff, pressing ENTER and ENTER again to turn all plots off.

**Follow these steps** to conduct a one-sample $t$-test.

1. Select STAT ► ►, accessing the TESTS menu. Select 2: T-Test, as shown in Figure 13.5, and press ENTER. Highlight Data and press ENTER to make it "stick". Use the down arrow key, ▼, to enter 98.6, the mean of the

null hypothesis. Use the down arrow key, ▼, to enter 2nd L1 in List:. Enter a frequency of 1 in Freq:. Use the down arrow key, ▼, to highlight the alternative hypothesis $< \mu_0$ and press ENTER to make it "stick". Use the down arrow key, highlighting Calculate and ENTER to execute the command, as shown in Figure 13.6. The results are shown in Figure 13.7.

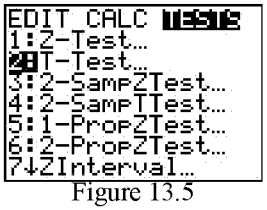


Figure 13.5

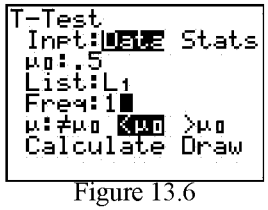


Figure 13.6

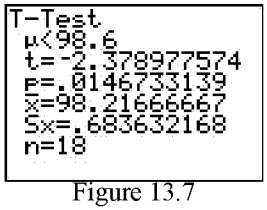


Figure 13.7

The TI output, shown in Figure 13.7, indicates the $Ha : \mu < 98.6$; the test statistic, $t = -2.38$; the $p$-value of the test statistic is $0.015$. The sample mean is $\overline{x} = 98.217$; the sample standard deviation is $s = 0.684$; the sample size is $n = 18$. Using a = 0.05as the level of significance criterion, the results are statistically significant because $0.015$, the p-value of the test, is less than $0.05$. In other words, we can reject the null hypothesis.
Wecan conclude, based on these data, that the mean temperature in the human population is actually less than 98.6 degrees.

The $p$-value represents the area to the left of $t = -2.38$ in a $t$-distribution with $n - 1 = 18 - 1 = 17$ degrees of freedom.

**Follow these steps** to obtain the $p$-value for this one-sample $t$-test.

1. Select **Follow these steps** to find the probability of observing a test statistic of $t = -2.38$, or less, given that the degrees of freedom is 17.

2. Calculate the probability of $t = -2.38$ or less.

   Press 2nd DISTR, located on row 4, column 4, above VARS ,to obtain the distribution function menu.

   a. Use the down arrow key, ▼, selecting 5: tcdf( , the Student's-t cumulative distribution probability function, as shown in Figure 13.8. Press ENTER. Enter the lowerbound, upperbound, and degrees of freedom, $df$. Type -1E99 , -2.38 , 17 ), as shown in Figure 13.9. Be sure to use the negation key found on the bottom row, column 4. The E is found on row six, column 2 above the , and is labeled as EE. Press ENTER to execute the command.

The output from the TI calculator is displayed in Figure 13.10.

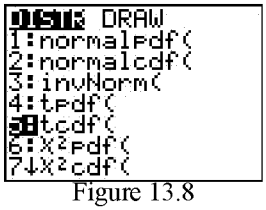


Figure 13.8

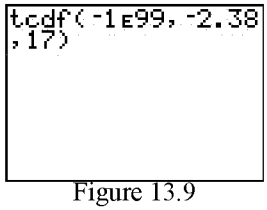


Figure 13.9

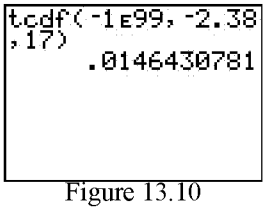


Figure 13.10

This TI calculator output, as shown in Figure 13.10, tells us that the probability of finding a value of $t = -2.38$ or less, given that the degrees of freedom is 17, is about 0.015 or 1.5% of the time.

## 13.3 Testing Hypotheses About Paired Differences

Remember from Chapter 11that the term paired data is used to describe data collected in natural pairs. In Chapter 4 we learned about *matched-pair* designs in which units are paired and one in each pair receives each treatment. Data are collected in pairs in other types of studies as well, such as when cases and controls are matched in case-control studies. Often, paired data occur when the researcher collects two measurements from each observational unit. For instance, if we record weights both before and after a diet program for each person in a sample, we have paired data. Or, we might record performance on a college entrance exam for each individual before and after a training program designed to boost performance on that type of exam.

In most cases paired data are collected because the researchers want to know about the differences, and not about the original observations. In particular, it is often of interest to know if the mean difference in the population is different from 0.A one-sample $t$-test can be used on the sample of differences to examine whether the sample mean difference is significantly different from 0. When this is done, the test is called a **paired $t$-test**. It is nothing more than a one sample $t$-test conducted on the $n$ differences. To emphasize that the data used in the test are differences, it is commonplace to use $d_i$ instead of $\overline{x}$ to denote the original data values, $\overline{d}$ instead of $\overline{x}$ for the sample mean of the differences, $s_d$ instead of $s$ for the sample standard deviation of the differences, and $\mu_d$ instead of $\mu$ for the population mean of the differences. Other than those notational differences, once the difference value has been computed for each pair, the test proceeds exactly like a one-sample $t$-test.

**Example 13.2 - The Effect of Alcohol on Useful Consciousness**

Ten pilots performed tasks at a simulated altitude of 25,000 feet. Each pilot performed the tasks in a completely sober condition and, three days later, after drinking alcohol. The response variable is the time (in seconds) of useful performance of the tasks for each condition. The longer a pilot spends on useful performance, the better. The research hypothesis is that useful performance time decreases with

alcohol use, so the data of interest is the decrease (or increase) in performance with alcohol compared to when sober. The data (Devore and Peck, 1993,p. 575) are as follows:

| Pilot | No Alcohol | Alcohol |
|---|---|---|
| 1 | 261 | 185 |
| 2 | 565 | 375 |
| 3 | 900 | 310 |
| 4 | 630 | 240 |
| 5 | 280 | 215 |
| 6 | 365 | 420 |
| 7 | 400 | 405 |
| 8 | 735 | 205 |
| 9 | 430 | 255 |
| 10 | 900 | 900 |

Table 13.2

This is a paired-data design. Let $\mu_d$ = population mean difference between no alcohol and alcohol measurements, if all pilots were to take these tests. Null and alternative hypotheses about $\mu_d$ are

$$H_0 : \mu_d = 0 \text{ seconds}$$
$$H_a : \mu_d > 0 \text{ seconds (i.e. no alcohol>alcohol)}$$

The alternative hypothesis is one-sided because we hope to show that if performance does change, there is longer useful performance in the "No Alcohol" condition.

**Follow these steps** to construct a modified boxplot checking the conditions required before conducting a paired $t$-test.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Turn off all StatPlots.

   Press [2nd] [STAT PLOT] accessing the StatPlot menu. Select 4: PlotsOff, pressing [ENTER] and [ENTER] again to turn all plots off.
   c. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

Press [STAT] [ENTER] to select the [STAT] list editor.

a. Enter the data found in Table 13.2.

Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the data for the No Alcohol group: 261, 565, 900, ... pressing [ENTER] after each entry, as shown in Figure 13.11.
Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the data for the Alcohol group: 185, 375, 310, ... pressing [ENTER] after each entry, as shown in Figure 13.11.

b. Obtain the paired differences.

Place the cursor on list L3, at the top and type [2nd] [L1] [-] [2nd] [L2], as shown in Figure 13.12. Press [ENTER] to obtain the paired differences, as shown in Figure 13.13.

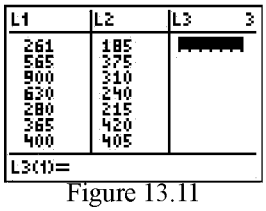


Figure 13.11

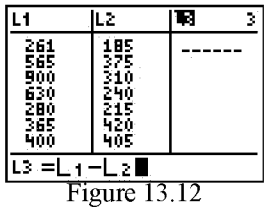


Figure 13.12

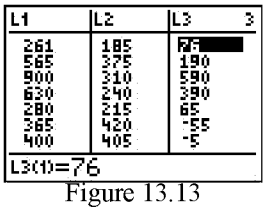


Figure 13.13

c. Plot the statistical data by creating a modified boxplot of the paired differences.

Press [2nd] [STAT PLOT] accessing the StatPlot menu, as shown in Figure 13.14. Press [ENTER], selecting Plot 1. Place the cursor on ON and press [ENTER]. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press [ENTER]. Use the down arrow key to select L3 as the list, [2nd] [L3], as shown in Figure 13.15. Enter a frequency of 1 in Freq:. Press [ZOOM] 9: ZoomStat to view the modified boxplot as shown in Figure 13.16. Use the [TRACE] key to obtain the five-number summary of the data.

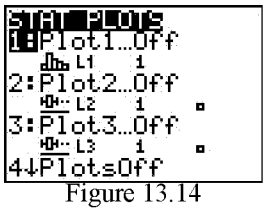


Figure 13.14

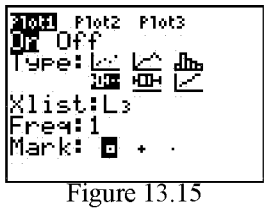


Figure 13.15

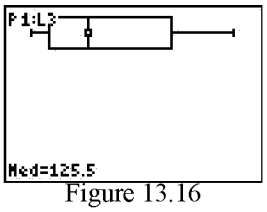


Figure 13.16

Figure 13.16 displays a boxplot of the differences for the ten participants. A

check of the necessary conditions for doing a $t$-test reveals that while the sample size is small and the dataset of differences for the sample does have some skewness, outliers and extreme skewness do not appear to be serious problems. So, a paired $t$-test will be used to examine this question.

**Follow these steps** to conduct a paired $t$-test.

1. Select [STAT] [►] [►], accessing the TESTS menu. Select 2: T-Test, as shown in Figure 13.17, and press [ENTER]. Highlight Data and press [ENTER] to make it "stick". Use the down arrow key, [▼], to enter 0.0, the mean of the null hypothesis. Use the down arrow key, [▼], to enter [2nd] [L3] in List:. Enter a frequency of 1 in Freq:. Use the down arrow key, [▼], to highlight the alternative hypothesis $> \mu_0$ and press [ENTER] to make it "stick". Use the down arrow key, highlighting Calculate and [ENTER] to execute the command, as shown in Figure 13.18. The results are shown in Figure 13.19.

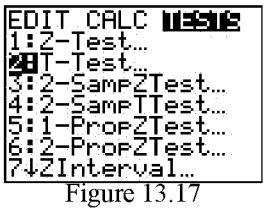


Figure 13.17

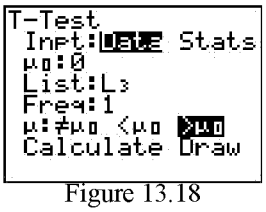


Figure 13.18

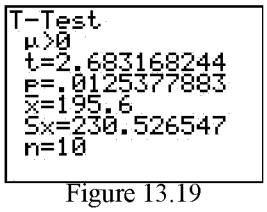


Figure 13.19

The TI output, shown in Figure 13.19, indicates the $Ha : \mu > 0$; the test statistic, $t = 2.68$; the $p$-value of the test statistic is 0.013. The sample mean is $\overline{x} = 195.6$; the sample standard deviation is $s = 230.5$; the sample size is $n = 10$. Using a = 0.05 as the level of significance criterion, the results are statistically significant because 0.013, the p-value of the test, is less than 0.05. In other words, we can reject the null hypothesis.
We can conclude, based on these data, that we can declare that alcohol has a statistically significant effect and decreases useful performance time.

## 13.4 Testing Hypotheses About Two Population Means

**The General (Unpooled) Case**
It is often of interest to determine whether the means of populations represented by two independent samples of a quantitative variable differ. The two populations may be represented by two categories of a categorical variable, such as males and females, or they may be two hypothetical populations represented by different treatment groups in an experiment. In most cases, when comparing two means, the null hypothesis is that they are equal:

$$H_0 : \mu_1 - \mu_2 = 0 \quad (\text{or } \mu_1 = \mu_2)$$

The procedure for testing this null hypothesis is called the **two-sample *t*-test**. Let's look at an example of a situation where a two-sample *t*-test would be used.

**Example 13.3 - The Effect of a Stare on Driving Behavior**

In Example 11.11, we discussed an experiment done by social psychologists at the University of California at Berkeley. The researchers either did not stare or did stare at automobile drivers stopped at a campus stop sign. In the experiment, the response variable was the time (in seconds) it took the drivers to drive from the stop sign to a mark on the other side of the intersection. The two populations represented by the observed times are the hypothetical ones that would consist of the the times drivers like these would take to move through a similar intersection, either under normal conditions (no stare) or the experimental condition (stare). An hypothesis the researchers wished to test was that the stare would speed up the crossing times, so the mean crossing time would be greater (slower) for those who did not experience the stare than it would be for those who did. The data are in the following table:

| NoStare | Stare |
|---|---|
| 8.3 | 5.6 |
| 5.5 | 5.0 |
| 6.0 | 5.7 |
| 8.1 | 6.3 |
| 8.8 | 6.5 |
| 7.5 | 5.8 |
| 7.8 | 4.5 |
| 7.1 | 6.1 |
| 5.7 | 4.8 |
| 6.5 | 4.9 |
| 4.7 | 4.5 |
| 6.9 | 7.2 |
| 5.2 | 5.8 |
| 4.7 | |

Table 13.3

So the null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \text{ or } H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 - \mu_2 > 0 \text{ or } H_a : \mu_1 > \mu_2$$

In these hypotheses, the subscript 1 is used to denote the No Stare population and the subscript 2 is used to denote the Stare population.

**Follow these steps** to construct modified boxplots checking the conditions required before conducting a two-sample $t$-test.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Turn off all StatPlots.

Press [2nd] [STAT PLOT] accessing the StatPlot menu. Select 4: PlotsOff, pressing [ENTER] and [ENTER] again to turn all plots off.

c. Clear all lists in the Stat editor.

Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

Press [STAT] [ENTER] to select the [STAT] list editor.

a. Enter the data found in Table 13.3.

Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the data for the No Stare group: 8.3, 5.5, 6.0 ... pressing [ENTER] after each entry, as shown in Figure 13.20.
Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the data for the Stare group: 5.6, 5.0, 5.7, ... pressing [ENTER] after each entry, as shown in Figure 13.20.

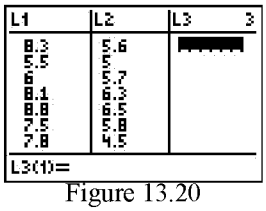


Figure 13.20

b. Plot the statistical data by creating modified boxplots of the crossing times.

Press [2nd] [STAT PLOT] accessing the StatPlot menu, as shown in Figure 13.21. Press [ENTER], selecting Plot 1. Place the cursor on ON and press [ENTER]. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press [ENTER]. Use the down arrow key to select L1 as the list, [2nd] [L1], as shown in Figure 13.22. Enter a frequency of 1 in Freq:.
Use the up arrow key, [▲], and the right arrow key, [►], placing the cursor on Plot2. Press [ENTER], selecting Plot 2. Place the cursor on ON and press [ENTER]. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press [ENTER]. Use the down arrow key to select L2 as the list, [2nd] [L2], as shown in Figure 13.23. Enter a frequency of 1 in Freq:.
Press [ZOOM] 9: ZoomStat to view the modified boxplot as shown in

Figure 13.24. Use the TRACE key to obtain the five-number summary of the data.

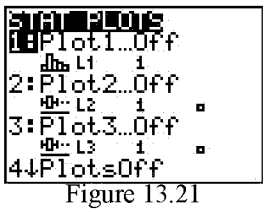


Figure 13.21

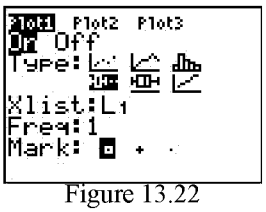


Figure 13.22

Figure 13.23

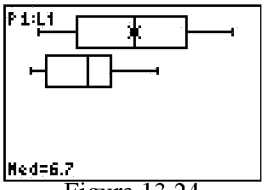


Figure 13.24

Figure 13.24 displays modified boxplots of the crossing times for the No Stare and Stare groups. A check of the necessary conditions for doing a $t$-test reveals that there were no extreme outliers and no extreme skewness in the data. So, a two-sample $t$-test will be used to test this null hypothesis.

**Follow these steps** to conduct an unpooled two-sample $t$-test.

1. Select STAT ► ►, accessing the TESTS menu. Select 4: 2-SampTTest, as shown in Figure 13.25, and press ENTER. Highlight Data and press ENTER to make it "stick". Use the down arrow key, ▼, to enter 2nd L1 in List1:. Use the down arrow key, ▼, to enter 2nd L2 in List2:. Enter a frequency of 1 in Freq1:. Enter a frequency of 1 in Freq:2. Use the down arrow key, ▼, to highlight the alternative hypothesis $> \mu_0$ and press ENTER to make it "stick". Use the down arrow key, ▼, highlighting No in Pooled: and press ENTER to make it "stick". Use the down arrow key, highlighting Calculate and ENTER to execute the command, as shown in Figure 13.26. The results are shown in Figure 13.27.

Figure 13.25

Figure 13.26

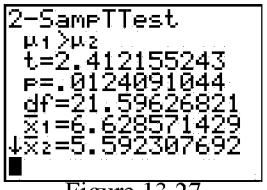


Figure 13.27

The TI output, shown in Figure 13.27, indicates the $Ha : \mu_1 > \mu_2$; the test statistic, $t = 2.41$; the $p$-value of the test statistic is $0.013$. The sample mean for the No Stare group is $\overline{x}_1 = 6.63$; the sample mean for the Stare group is $\overline{x}_2 = 5.59$. Using the down arrow key, $\boxed{\blacktriangledown}$, reveals the the sample standard deviation for the No Stare group is $s_1 = 1.36$ and for the Stare group is $s_2 = 0.822$. The sample size for the No Stare group is $n_1 = 14$; the sample size for the Stare group is $n_1 = 13$. Using a = 0.05 as the level of significance criterion, the results are statistically significant because $0.013$, the p-value of the test, is less than $0.05$. In other words, we can reject the null hypothesis.
We can conclude that if all drivers were stared at, the mean crossing times at an intersection would be faster than under normal condions. It can also be noted that these data were collected in a randomized experiment, so it's reasonable to conclude that the staring or not staring caused the difference between the mean crossing times.

**The Pooled Two-Sample t-Test**
The use of the *t*-distribution for finding p-values for the two-sample *t*-test is an approximation. However, if the population standard deviations are the same, then there is a more precise method available. The **pooled two-sample *t*-test** is based on the assumption that $\sigma_1 = \sigma_2 = \sigma$, the common standard deviation for both populations. The sample variances are combined to provide a **pooled sample variance,** $s_p^2$, and **pooled standard deviation**, $s_p$. The formula for the pooled sample variance is

$$\text{Pooled sample variance} = s_p^2 = \frac{(n_1 - 1)\, s_1^2 \; + \; (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

The computations for the *t*-statistic and *p*-value remain the same as with the regular two-sample *t*-test except for two small changes. First, the pooled sample variance is used in place of the individual sample variances when computing the standard error. Second, the degrees of freedom are $df = n_1 + n_2 - 2$.

**Example 13.6 - Legitimate Pooled t-Test for Comparing Male and Female Sleep Time**

In Example 11.14, we found a confidence interval for the difference in mean sleep time for male and female college students. In that example, students in an introductory statistics class filled out a survey on a variety of issues, including how much sleep they had the previous night. Let's use the data to estimate how much more or less male students sleep than female students, on average. The class included 83 females and 65 males. Let's assume these students are equivalent to a random sample of all students who take introductory statistics. Is there a significant difference between how long female and male students represented by this sample slept the previous night? To answer this question, we'll conduct a two-sample $t$-test involving the null hypothesis $\mu_1 - \mu_2 = 0$ = difference in mean sleep hours for females

versus males, involving both the unpooled and pooled versions. Here is the data.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3.50 | 4.00 | 5.00 | 8.50 | 7.00 | 6.50 | 5.00 | 9.00 | 9.00 |
| 6.50 | 5.50 | 7.00 | 9.50 | 7.00 | 6.50 | 5.00 | 6.00 | 7.00 |
| 7.00 | 5.50 | 7.50 | 9.00 | 7.00 | 6.00 | 8.50 | 7.00 | 11.00 |
| 8.00 | 4.00 | 7.00 | 4.50 | 4.00 | 7.00 | 9.50 | 9.00 | 9.00 |
| 8.00 | 8.50 | 6.50 | 5.00 | 5.25 | 7.50 | 8.00 | 6.50 | 6.00 |
| 10.00 | 7.00 | 8.00 | 8.50 | 6.00 | 7.00 | 4.00 | 7.00 | 10.00 |
| 8.00 | 7.00 | 6.00 | 10.00 | 4.50 | 6.00 | 3.00 | 5.00 | 4.50 |
| 7.00 | 9.00 | 8.00 | 8.00 | 8.00 | 6.50 | 6.00 | 10.00 | 8.00 |
| 9.00 | 7.00 | 8.00 | 4.50 | 8.00 | 9.50 | 8.00 | 8.00 | 8.00 |
| 5.00 | 6.00 | | | | | | | |

Table 13.4 Sleep Hours for Females

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9.0 | 6.0 | 6.0 | 9.0 | 8.0 | 4.5 | 8.0 | 8.0 | 5.0 |
| 5.0 | 4.0 | 8.5 | 5.0 | 8.5 | 9.0 | 6.0 | 6.5 | 5.0 |
| 6.0 | 7.0 | 8.0 | 7.0 | 6.5 | 6.0 | 6.5 | 9.0 | 4.0 |
| 3.5 | 4.5 | 5.0 | 8.0 | 7.5 | 8.0 | 2.0 | 9.0 | 7.5 |
| 12.0 | 8.0 | 5.0 | 5.0 | 6.5 | 6.5 | 7.0 | 7.0 | 7.0 |
| 7.0 | 6.0 | 8.0 | 5.0 | 6.0 | 6.0 | 7.5 | 5.5 | 7.0 |
| 9.0 | 7.0 | 6.0 | 6.0 | 5.5 | 7.0 | 5.5 | 6.0 | 4.0 |
| 7.0 | 5.0 | | | | | | | |

Table 13.5 Sleep Hours for Males

**Follow these steps** to construct two modified boxplots checking the conditions required for using the two-sample $t$-test.

1. Preparations:
   a. Turn off all"Y=" functions.

      Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Turn off all StatPlots.

      Press [2nd] [STAT PLOT] accessing the StatPlot menu. Select 4: PlotsOff, pressing [ENTER] and [ENTER] again to turn all plots off.
   c. Clear all lists in the Stat editor.

      Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

a. Enter the data found in Table 13.4 and in Table 13.5.

Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the data for Female Sleep Hours: 3.5, 4.0, 5.0 ... pressing ENTER after each entry, as shown in Figure 13.28.
Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the data for Male Sleep Hours:9.0, 6.0, 6.0 ... pressing ENTER after each entry, as shown in Figure 13.28.

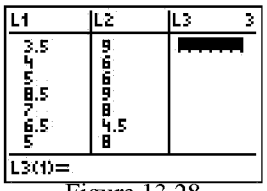


Figure 13.28

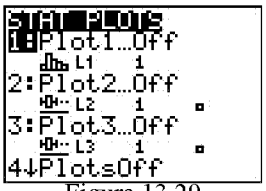


Figure 13.29

Figure 13.30

So the null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \text{ or } H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 - \mu_2 \neq 0 \text{ or } H_a : \mu_1 \neq \mu_2$$

In these hypotheses, the subscript 1 is used to denote the female population and the subscript 2 is used to denote the male population.

b. Plot the statistical data by creating two modified boxplots of the sleep times for females and males.

Press 2nd STAT PLOT accessing the StatPlot menu, as shown in Figure 13.29. Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select L1 as the list, 2nd L1, as shown in Figure 13.30. Enter a frequency of 1 in Freq:. Use the up arrow ▲, placing the cursor on Plot 2. Press ENTER, selecting Plot 2. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select L2 as the list, 2nd L2, as shown in Figure 13.31. Enter a frequency of 1 in Freq:. Press ZOOM 9: ZoomStat to view the two modified boxplots of the sleep times for females and males as shown in Figure 13.32. Use the TRACE key to obtain the

five-number summaries of the data.

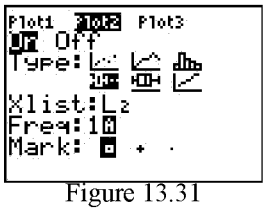


Figure 13.31

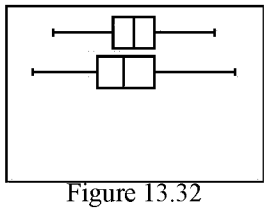


Figure 13.32

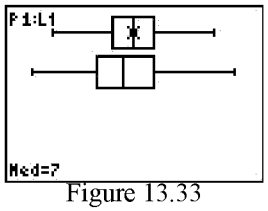


Figure 13.33

Figure 13.33 shows the two modified boxplots of the sleep times for females and males. The modified boxplots suggest that there are no major outliers or extreme skewness. The sample median for females is 7.0 and the sample median for males is 5.0 hours, respectively. In data ranging from 3 to 11 for females and from 2 to 12 for males, these statistics are close enough to rule out extreme skewness. The two modified boxplots of the sleep times for females and males in Figure 13.33 provides further evidence that the appropriate conditions are satisfied.

**Follow these steps** to conduct an unpooled and a pooled two-sample $t$-test.

1. Conduct the unpooled two-sample $t$-test.
   Select [STAT] [►] [►], accessing the TESTS menu. Select 4: 2-SampTTest, as shown in Figure 13.34, and press [ENTER]. Highlight Data and press [ENTER] to make it "stick". Use the down arrow key, [▼], to enter [2nd] [L1] in List1:. Use the down arrow key, [▼], to enter [2nd] [L2] in List2:. Enter a frequency of 1 in Freq1:. Enter a frequency of 1 in Freq:2. Use the down arrow key, [▼], to highlight the alternative hypothesis $> \mu_0$ and press [ENTER] to make it "stick". Use the down arrow key, [▼], highlighting No in Pooled: and press [ENTER] to make it "stick". Use the down arrow key, highlighting Calculate and [ENTER] to execute the command, as shown in Figure 13.35. The results are shown in Figure 13.36.

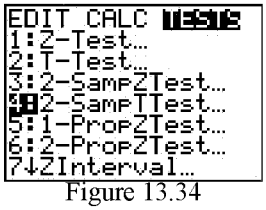


Figure 13.34

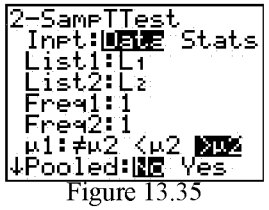


Figure 13.35

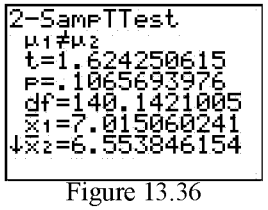


Figure 13.36

The TI output, shown in Figure 13.36, indicates the $Ha : \mu_1 \neq \mu_2$; the test statistic, $t = 1.62$; the $p$-value of the test statistic is 0.11. The sample mean for the Female Sleep Times is $\overline{x}_1 = 7.02$; the sample mean for the Male Sleep Times is $\overline{x}_2 = 6.55$. Using the down arrow key, [▼], reveals the the sample standard deviation for the Female Sleep Times is $s_1 = 1.75$ and for the Male Sleep Times is $s_2 = 1.68$.

The sample size for the Female Sleep Times is $n_1 = 83$; the sample size for the Male Sleep Times is $n_2 = 65$. Using a = 0.05 as the level of significance criterion, the results are not statistically significant because 0.11, the p-value of the test, is more than 0.05. In other words, we can not reject the null hypothesis.

1. Conduct the pooled two-sample $t$-test.
   Select [STAT] [►] [►], accessing the TESTS menu. Select 4: 2-SampTTest, as shown in Figure 13.37, and press [ENTER]. Highlight Data and press [ENTER] to make it "stick". Use the down arrow key, [▼], to enter [2nd] [L1] in List1:. Use the down arrow key, [▼], to enter [2nd] [L2] in List2:. Enter a frequency of 1 in Freq1:. Enter a frequency of 1 in Freq:2. Use the down arrow key, [▼], to highlight the alternative hypothesis $> \mu_0$ and press [ENTER] to make it "stick". Use the down arrow key, [▼], highlighting Yes in Pooled: and press [ENTER] to make it "stick". Use the down arrow key, highlighting Calculate and [ENTER] to execute the command, as shown in Figure 13.38. The results are shown in Figure 13.39.

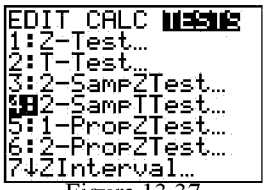


Figure 13.37

Figure 13.38

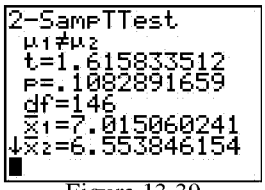


Figure 13.39

The TI output, shown in Figure 13.39, indicates the $Ha : \mu_1 \neq \mu_2$; the test statistic, $t = 1.62$; the $p$-value of the test statistic is 0.11. The sample mean for the Female Sleep Times is $\overline{x}_1 = 7.02$; the sample mean for the Male Sleep Times is $\overline{x}_2 = 6.55$. Using the down arrow key, [▼], reveals the the sample standard deviation for the Female Sleep Times is $s_1 = 1.75$ and for the Male Sleep Times is $s_2 = 1.68$. Furthermore the pooled standard deviation is $s_{xp} = 1.72$. The sample size for the Female Sleep Times is $n_1 = 83$; the sample size for the Male Sleep Times is $n_2 = 65$.
The TI output, shown in Figure ??.?, indicates the sample sizes are similar, as are the sample standard deviations. Consequently, the pooled standard deviation of 1.72 is similiar to the seperate standard deviations of 1.75 and 1.68. The test statistic of 1.62 is identical under both procedures, and the $p$-value of 0.11 is the same. The use of the pooled procedure was warranted in this case.

# Chapter 14
# Inference About Simple Regression

We learned in Chapter 5 that often a straight line describes the pattern of a relationship between two quantitative variables. For instance, in Example 5.1 we explored the relationship between the handspans (cm) and heights (in.) of 167 college students and found that the pattern of the relationship in this sample could be described by the equation
Average handspan = -3 + 0.35 Height
An equation like the one relating handspan to height is called a regression equation, and the term simple regression is sometimes used to describe the analysis of a straight-line relationship (linear relationship) between a response variable (y variable) and an explanatory variable (x variable).
In Chapter 5, we only used regression methods to describe a sample and did not make statistical inferences about the larger population. Now, we consider how to make inferences about a relationship in the population represented by the sample. Some questions involving the population that we might ask when analyzing a relationship are

1. Does the observed relationship also occur in the population? For example, is the observed relationship between handspan and height strong enough to conclude that the relationship also holds in the population?
2. For a linear relationship, what is the slope of the regression line in the population? For example, in the larger population, what is the slope of the regression line that connects handspans to heights?
3. What is the mean value of the response variable (y) for individuals with a specific value of the explanatory variable (x)? For example, what is the mean handspan in a population of people 65 inches tall?
4. What interval of values predicts the value of the response variable (y) for an individual with a specific value of the explanatory variable (x)? For example, what interval predicts.the handspan-of-an individual 65 inches tall?

After reading this chapter you should be able to:

1. Obtain the regression output, identifying the slope, intercept, $r^2$, SSTO, and SSE for two quantitative variables.
2. Calculate, $s$, the stand deviation for regression.
3. Test a hypotheses about a population slope.

4. Test a hypotheses about a population correlation coefficient.
5. Find a prediction interval for the response variable $(y)$ for an individual with a particular value of $x$.
6. Find a confidence interval for the mean value of $y$ for a population of individuals who all have the same particular value of $x$.
7. Create residual plots.
8. Perfrom a natural log transformation of variables.

**Keystrokes Introduced**

1. [STAT] [►] to obtain the [STAT] CALC menu, selecting 8: LinReg (a+bx). The function 8: LinReg (a+bx) [*Xlistname*,Ylistname,freqlist,regequ] fits a linear regression equation to *Xlistname* and *Ylistname* with frequency *freqlist,* and stores the regression equation to *regequ.*
2. [2nd] [CATALOG], locating **DiagnosticOn**, to display $r$ and $r^2$ as regression model results.
3. [2nd] [LIST] [►] [►] selecting 5: sum(. **sum(***list* [,*start*,*end* ]) returns the sum of elements of *list* from *start* to *end.*
4. [2nd] [ANS] copies the last answer to the variable location.
5. [VARS], selecting 5: Statistics accessing the XY, $\sum$, EQ, and PTS variables. You will access the XY, $\sum$, and EQ variable menus in this chapter.
6. [VARS] [►] displays the Y-VARS menu, selects the 1:Function menu. You will use the Y-function Y1 in this chapter.
7. [STAT] [►] [►] to obtain the [STAT] TEST menu, selecting E: Linear RegTTest. **Linear RegTTest** performs a linear regression and a $t$-test about the population slope.
8. [STO→] [x,T,$\theta$,$n$] [ENTER] to store a number in the x register.
9. [LN] returns the natural logarithm of a real number or list.
10. [2nd] [$e^x$] returns $e$ raised to a power or a list.

## 14.1 Sample and Population Regression Models

A regression model describes the relationship between a quantitative response variable (the y variable) and one or more explanatory variables (x variables). The y variable is sometimes called the *dependent variable*, and because regression models may be used to make predictions, the x variables may be called the *predictor variables*.The labels *response variable* and *explanatory variable* may be used for the variables on the y axis and x axis, respectively, even if there is not an obvious way to assign these labels in the usual sense.

Any regression model has two important components. The most obvious com-

ponent is the equation that describes how the mean value of the y variable is connected to specific values of the x variable. The equation for the connection between handspan and height, Average handspan = -3 + 0.35 (Height), is an example. In this chapter, we focus on *linear relationships*, so a straight-line equation will be used, but it is important to note that some relationships are *curvilinear*.
The second component of a regression model describes how individuals vary from the regression line. Figure 14.1, in the text, which is identical to Figure 5.6 in the text, displays the raw data for the sample of n = 167 handspans and heights along with the regression line that estimates how the mean handspan is connected to specific heights. Notice that most individuals vary from the line. When we examine sample data, we will find it useful to estimate the general size of the deviations from the line. When we consider a model for the relationship within the population represented by a sample, we will state assumptions about the distribution of deviations from the line.

### The Regression Line for the Sample

In Chapter 5, we introduced this notation for the regression line that describes sample data:

$$\widehat{y} = b_0 + b_1 x$$

In any given situation, the sample is used to determine values for $b_0$ and $b_1$.
$\widehat{y}$ is spoken as "y-hat" and it is also referred to either as predicted y or estimated y.
$b_0$ is the intercept of the straight line. The intercept is the value of y when $x = 0$.
$b_1$ is the slope of the straight line. The slope tells us how much of an increase (or decrease) there is for y when the x variable increases by one unit. The sign of the slope tells us whether y increases or decreases when x increases. If the slope is 0, there is no linear relationship between x and y because y is the same for all values of x.

## 14.2 Estimating the Standard Deviation for Regression

Recall that the standard deviation in the regression model measures, roughly, the average deviation ofyvalues from the mean (the regression line). Expressed another way, the **standard deviation for regression** measures the general size of the residuals. This is an important and useful statistic for describing individual variation in a regression problem, and it also provides information about how accurately the regression equation might predict y values for individuals. A relatively small standard deviation from the regression line indicates that individual data points generally fall close to the line, so predictions based on the line will be close to the actual values.
The calculation of the estimate of standard deviation is based on the sum of the squared residuals for the sample. This quantity is called the **sum of squared errors** and is denoted by **SSE**. Synonyms for "sum of squared errors" are **residual sum of squares** or **sum of squared residuals**. To find the SSE, residuals are calcu-

lated for all observations, then the residuals are squared and summed. The *standard deviation from the regression line* is

$$s = \frac{Sum\ of\ Squared\ \text{Residuals}}{n-2} = \sqrt{\frac{SSE}{n-2}}$$

and this sample statistic estimates the population standard deviation $\sigma$.

## The Proportion of Variation Explained by x

In Chapter 5, we learned that a statistic denoted as $r^2$ is used to measure how well the explanatory variable actually does explain the variation in the response variable. This statistic is also denoted as $R^2$ (rather than $r^2$), and the value is commonly expressed as a percent. Researchers typically use the phrase "proportion of variation explained by x" in conjunction with the value of $r^2$.For example, if $r = 0.60$ (or 60%), the researcher may write that the explanatory variable explains 60%of the variation in the response variable.
The formula for $r^2$ presented in Chapter 5 was

$$r^2 = \frac{SSTO-SSE}{SSTO}$$

The quantity SSTO is the sum of squared differences between observed y values and the sample mean y. It measures the size of the deviations of the y values from the overall mean of y, whereas SSE measures the deviations of the y values from the predicted values y.

**Example 14.5 - Driver Age and Highway Sign-Reading Distance**

In a study of the legibility and visibility of highway signs, a Pennsylvania research firm determined the maximum distance at which each of 30 drivers could read a newly designed sign. The 30 participants in the study ranged in age from 18 to 82 years old. The government agency that funded the research hoped to improve highway safety for older drivers and wanted to examine the relationship between age and the sign legibility distance.
Table 14.1 lists the data; Table 5.4 also lists the data. We will use the TI calculator to determine the linear relationship between "maximum distance" and "age", identifying the slope, intercept, $r^2$, SSTO, and SSE for this set of measurements.

| Age | Distance | Age | Distance | Age | Distance |
|---|---|---|---|---|---|
| 18 | 510 | 37 | 420 | 68 | 300 |
| 20 | 590 | 41 | 460 | 70 | 390 |
| 22 | 560 | 46 | 450 | 71 | 320 |
| 23 | 510 | 49 | 380 | 72 | 370 |
| 23 | 460 | 53 | 460 | 73 | 280 |
| 25 | 490 | 55 | 420 | 74 | 420 |
| 27 | 560 | 63 | 350 | 75 | 460 |
| 28 | 510 | 65 | 420 | 77 | 360 |
| 29 | 460 | 66 | 300 | 79 | 310 |
| 32 | 410 | 67 | 410 | 82 | 360 |

Table 14.1

**Follow these steps** to obtain the regression output, identifying the slope, intercept, $r^2$, SSTO, and SSE for this set of measurements.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

   c. Turn Diagnostics On to display $r$, the correlation coefficient, and $r^2$, the coefficient of determination.

   Press [2nd] [CATALOG], located on the bottom row, 2nd column from the left above the [0]. Press [ALPHA] [D], and use the down arrow key to locate **DiagnosticOn**. Press [ENTER] to select the command and press [ENTER] once again to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the "Age" data: 18, 20, 22, ... pressing [ENTER] after each entry.
   Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the "Distance" data: 510, 590, 560, ... in L2 pressing [ENTER] after each entry, as shown in Figure 14.1.

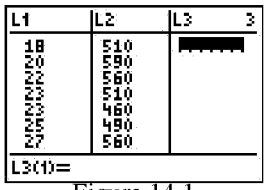



Figure 14.1

3. Obtain the regression equation.

   Press [STAT] [▶] to obtain the [STAT] CALC menu.

   a. Use the down arrow key, [▼], seven times and press [ENTER], or just press [8] to select 8: LinReg (a+bx), as shown in Figure 14.2. Press [2nd] [L1] to select

the "Age" data. Press , 2nd L2 to select the "Distance" data, as shown in Figure 14.3. Press ENTER to execute the command. The output from the TI calculator is displayed in Figure 14.4.

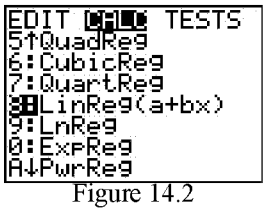


Figure 14.2

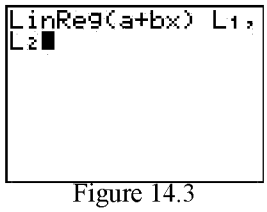


Figure 14.3

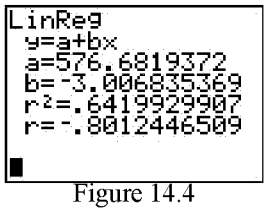


Figure 14.4

The sample regression equation is $\widehat{y} = 577 - 3.01x$; the $y$-intercept is 577 and the slope is $-3.01$. The correlation coefficient, $r = 0.801$ describes a moderately strong positive association. The squared correlation is $r = (0.801)^2 = 0.642$. The "$x$ = Age" explains 64.2% of the variation among the "$y$ =Distance" data.

4. Obtain the sum of square errors and the standard deviation for regression.

Press VARS, row 4, column 4. Select 5: Statistics. Use the right arrow, ▶, twice, selecting 2: a. Press ⊞, VARS, 5: Statistics, and the right arrow, ▶, twice, selecting 3: b. Press ×, 2nd L1. Press STO→ 2nd L3. Your screen should look like Figure 14.5. These data points represent the predicted values, $\widehat{y}$, from "$x$ =Age" variable stored in list L1. These predicted values, $\widehat{y}$, of "Distance" are stored in L3.

Press STAT ENTER to select the STAT list editor.

a. Place the cursor at the top of list L4. Press 2nd L2, - 2nd L3, as shown in Figure 14.6, pressing ENTER to obtain the residuals. The residuals are shown in Figure 14.6.

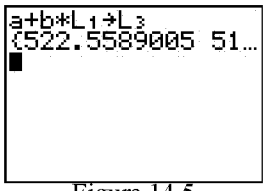


Figure 14.5

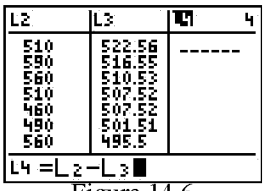


Figure 14.6

b. Press 2nd QUIT. Press 2nd LIST ▶ ▶, selecting 5: sum(. Press 2nd L4 $x^2$ ). Press ENTER. Your screen should look like Figure

14.8.

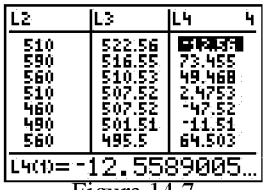


Figure 14.7

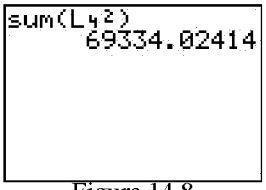


Figure 14.8

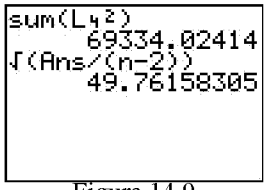


Figure 14.9

c. Use the equation, $s = \sqrt{\frac{SSE}{n-2}}$ to obtain the standard deviation for regression.

Press 2nd $\sqrt{}$, 2nd ANS ÷ ( VARS 5 1 − 2 ) and press ENTER. The results are shown in Figure 14.9.

The sum of squared errors is SSE = 69334. The value of s, the standard deviation for regression is 49.76 as shown in Figure 14.9

5. Obtain the total sum of squares, $SSTO = sum\ ((y\text{-}\overline{y})^2)$.

To obtain the mean of the predicted $y$ values, press STAT > CALC, selecting 1: 1Var Stats. Press ENTER. Press 2nd L3 and ENTER, as shown in Figure 14.10. The results are shown in Figure 14.11.

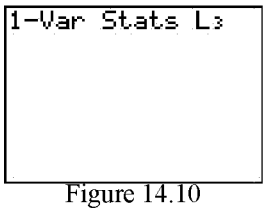


Figure 14.10

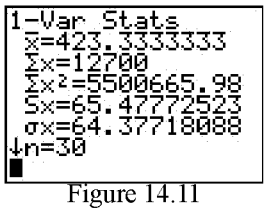


Figure 14.11

The mean of the predicted $y$ values is 423.33.

Press STAT ENTER to select the STAT list editor.

a. Place the cursor at the top of list L5. Press 2nd L2, - VARS ,selecting 5: Statistics. Select 2: $\overline{x}$ as shown in Figure 14.12, pressing ENTER. The results are shown in Figure 14.13.

b. Press 2nd QUIT. Press 2nd LIST ► ►, selecting 5: sum(. Press 2nd L5 $x^2$ ). Press ENTER. Your screen should look like Figure 14.14.

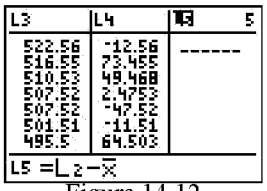


Figure 14.12

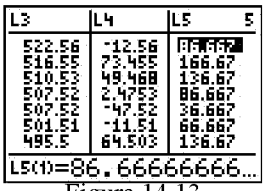


Figure 14.13

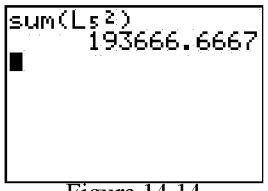


Figure 14.14

The total sum of quares, $SSTO = sum\ ((y\text{-}\overline{y})^2)$ is 193667.

The coefficient of determination, $r^2 = \frac{SSTO-SSE}{SSTO} = \frac{193667-69334}{193667} = 0.64199$.

## 14.3 Inference About the Linear Regression Relationship

When researchers do a regression analysis, they occasionally know based on past research or common sense that the variables are indeed related. In some instances, however, it may be necessary to do a hypothesis test in order to make the generalization that two variables are related in the population represented by the sample. The **statistical significance of a linear relationship** can be evaluated by testing whether or not the slope is 0.Recall that if the slope is 0 in a simple regression model, the two variables are not related because changes in the $x$ variable will not lead to changes in the $y$ variable. The usual null hypothesis and alternative hypothesis about $\beta_1$, the slope of the population regression line $E(Y) = \beta_0 + \beta_1 x$, are
$H_o : \beta_1 = 0$ (the population slope is 0, so y and x are not linearly related)
$H_1 : \beta_1 \neq 0$ : (the population slope is not 0, so y and x are linearly related)
The alternative hypothesis may be one-sided or two-sided, although most statistical software uses the two-sided alternative.
The test statistic used to do the hypothesis test is a t-statistic with the same general format used in Chapters 12 and 13.That format, and its application to this situation, is

$$t = \frac{\text{Sample Statistic-Null value}}{\text{Standard error}} = \frac{b_1 - 0}{\text{s.e.}(b_1)}$$

This is a standardized statistic for the difference between the sample slope and 0, the null value. Notice that a large value of the sample slope (either positive or negative) relative to its standard error will give a large value of t. If the mathematical assumptions about the population model described in Section 14.1 are correct, the statistic has a t-distribution with n $-$ 2 degrees of freedom. The p-value for the test is determined using that distribution.
"By hand" calculations of the sample slope and its standard error are cumbersome. Fortunately, the regression analysis of most statistical software includes a t-statistic and a p-value for this significance test.

**Example 14.5 - Continued - Driver Age and Highway Sign-Reading Distance**

In Example 14.3, you produced the output for the regression of sign-reading distance and driver age using your TI calculator. The sample estimate of the slope is $b_l = -3.01$. This sample slope is different from 0, but is it different enough to enable us to generalize that a linear relationship exists in the population represented by this sample?

**Follow these steps** to test hypotheses about the population slope.

1. Perform a linear regression t-test concerning the population slope:

   Press [STAT] [►] [►] to obtain the [STAT] TEST menu.

   a. Use the down arrow key, [▼] repeatedly to select E: Linear RegTTest, as shown in Figure 14.15 and press [ENTER]. Place [2nd] [L1] to place list L1 in the Xlist: position. Place [2nd] [L2] to place list L2 in the Ylist: position. Let the Freq: be 1. Select [≠ 0] for $\beta$ & $\rho$. Place the functionY1 in REgEq: by selecting [VARS] [►] [1] [ENTER] [1] [ENTER]. These settings are shown in Figure 14.16. Highlight Calculate and press [ENTER]. The results of the linear regression t-test concerning the population slope are shown in Figures 14.17 and 14.18.

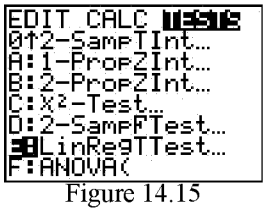


Figure 14.15

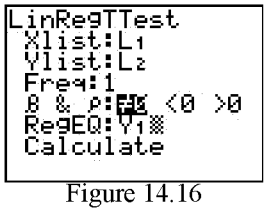


Figure 14.16

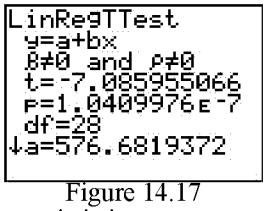


Figure 14.17

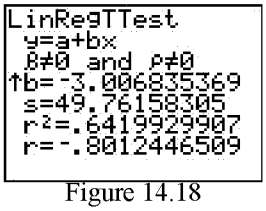


Figure 14.18

   The test statistic is

$$t = \frac{\text{Sample Statistic-Null value}}{\text{Standard error}} = \frac{b_1 - 0}{\text{s.e.}(b_1)} = -7.086$$

   The p-value is, to three decimal places, 0.000.This means the probability is virtually 0 that the observed slope could be as far from 0 or farther than it is if there is no linear relationship in the population. So, as we might expect for these variables, we can conclude that the relationship between the two variables in the sample represents a real relationship in the population. Note that the hypotheses tests for the population slope($H_0 : \beta_1 = 0$) and the hypotheses test for the correlation coefficient, $H_0 : \rho = 0$, are both done at the same time on the TI calculator.

## Confidence Interval for the Population Slope

The significance test of whether or not the population slope is 0 only tells us if we can declare the relationship to be statistically significant. If we decide that the true slope is not 0, we might ask, "What is the value of the slope?" We can answer this question with a confidence interval for$\beta_1$ the population slope.
The format for this confidence interval is the same as the general format used in

Chapters 10and 11, which is

Sample statistic±Multiplier×Standard error

The sample statistic is $b_1$, the slope of the least-squares regression line for the sample. As shown already, the standard error formula is complicated and we'll usually rely on statistical software to determine this value. The "multiplier" will be labeled t* and is determined using a t-distribution with $df = n - 2$. TableA.2 can be used to find the multiplier for the desired confidence level.
A confidence interval for $\beta_1$ is

$$b_1 \pm t^* s.e.(b_1)$$

The "multiplier" $t^*$is found using a t-distribution with $df = n - 2$ and is such that the probability between $-t^*$and $+t^*$ equals the confidence level for the interval.
Recall that

$$s.e.(b_1) = \frac{s}{\sqrt{\sum (x_i - x)^2}}, \text{ where } \sqrt{\frac{SSE}{n-2}}$$

**Follow these steps** to find the confidence interval for $\beta_1$.

1. Find the $s.e.(b_1)$.

   Find the sum of squares of x, $\sum (x_i - x)^2$.

   Press [STAT] > CALC, selecting 1: 1Var Stats. Press [ENTER]. Press [2nd] [L1] and [ENTER], as shown in Figure 14.19. The results are shown in Figure 14.20.


Figure 4.19

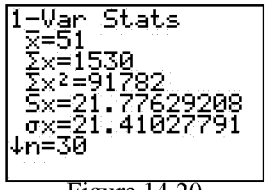



Figure 14.20

   To enter the equation for the sum of squares of x, $\sum (x_i - x)^2 = \sum x^2 - \frac{(\sum x)^2}{n}$, press [VARS] [5] [▶] [2] [-] [(] [VARS] [5] [▶] [1] [)] [$x^2$] [÷] [VARS] [5] [1] [ENTER].
   The sum of squares of x, $\sum (x_i - x)^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 91782 - \frac{(1530)^2}{30} = 13752$. The results are shown in Figure 14.21. Recall that we had previously found $s = \sqrt{\frac{SSE}{n-2}} = 49.76$. Enter $\frac{49.76}{\sqrt{13752}}$ to find $s.e.(b_1)$.

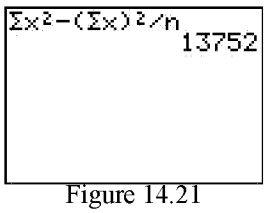



Figure 14.21

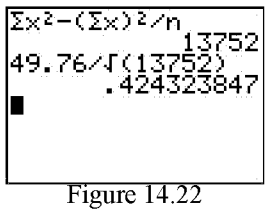



Figure 14.22

   The $s.e.(b_1) = \frac{s}{\sqrt{\sum (x_i - x)^2}} = \frac{49.76}{\sqrt{13752}} = 0.42432$. The results are shown in Figure 14.22.

2. Calculate the 95% confidence interval. The "multiplier" $t^*_{(\frac{\alpha}{2}=0.05, df=28)} = 2.05$ is found using TableA.2 in the text.

   Place $-3.01 + 2.05 \times 0.4243$ on the homescreen. Press ENTER.
   Place $-3.01 - 2.05 \times 0.4243$ on the homescreen. Press ENTER.

   The results of these calculations are shown in Figure 14.23.

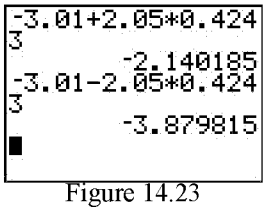



Figure 14.23

   With 95%confidence, we can estimate that in the population of drivers represented by this sample, the mean sign-reading distance decreases somewhere between 3.88 and 2.14 feet for each one-year increase in age.

## 14.4 Predicting $y$ and Estimating Mean $y$ at a Specific $x$

In this section we cover two different types of intervals used to make inferences about the response variable ($y$). The first type of interval *predicts the value of* $y$ for an individual with a specific value of $x$. For example, we may want to predict the freshman year GPA of a college applicant who has a 3.6 high school GPA. The second type of interval *estimates the mean value of* $y$ for a population of individuals who all have the same specific value of $x$. As an example, we may want to estimate the mean (average) of freshman year GPA of all college applicants who have a 3.6 high school GPA.

### Predicting the Value of y for an Individual

An important use of a regression equation is to estimate or predict the unknown value of a response variable for an individual with a known specific value of the explanatory variable. The formula for the prediction interval for $y$ at a specific $x$ is

$$\widehat{y} \pm t^* \sqrt{s^2 + [s.e.(fit)]^2}$$

where

$$s.e.(fit) = s\sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum(x_i - \overline{x})2}}$$

The "multiplier" $t^*$ is found using a t-distribution with $df = n - 2$. TableA.2 can be used to find the multiplier for the desired level for the interval..

**Example 14.5 - Continued - Driver Age and Highway Sign-Reading Distance**

Previously in Example 14.5, you tested a hypotheses about the population slope

for the regression of sign-reading distance and driver age using your TI calculator. You placed the function Y1 in REgEq: in the linear regression t-test screen. That function Y1 now stores the sample regression equation: $\widehat{y} = 577 - 3.01x$.

**Follow these steps** to predict the maximum distance at which an individual can read a highway sign by substituting his or her age for x in the sample regression equation.

1. Place a value of x in the sample regression equation.

   Press [2] [1] [STO→] [x,T,θ,n] [ENTER] to store 21 in the x register, as shown in Figure 14.24.

2. Evaluate sample regression equation: $\widehat{y} = 577 - 3.01x$, where x is replaced with 21.

   Press [VARS] [▶] [1] [1] [ENTER] to place the function Y1 on the homescreen. Press [ENTER] to evaluate Y1. The results are shown in Figure 14.25.

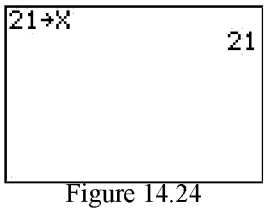


Figure 14.24

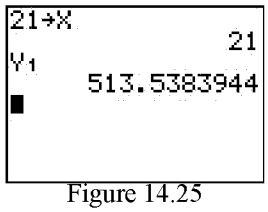


Figure 14.25

   For a person 21 years old, the predicted distance is approximately $\widehat{y}$ = 577 - 3(21) = 514 feet.

**Follow these steps** to obtain a 95% prediction interval of the maximum distance at which a 21 year-old individual can read a highway sign.

1. Find $s.e.(fit) = s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})2}}$. Recall $s = \sqrt{\frac{SSE}{n-2}} = 49.76$ and $\sum(x_i - x)^2 = 13752$ were found previously. In addition $n = 30$.

   Press 49.76 [*] [2nd] [$\sqrt{}$] [(] [(] [1] [÷] [30] [)] [+] [(] 21 [VARS] [5] [2] [)] [$x^2$] [÷] 13752 [)] and press [ENTER] to execute the command.The expression and results are shown in Figure 14.26.

2. Obtain the 95% prediction interval of the maximum distance at which a 21 year-old individual can read a highway sign.

   Press [VARS] [▶] [1] [1] [ENTER] to place the function Y1 on the homescreen. Recall that the "multiplier" $t^*_{(\frac{\alpha}{2}=0.05, df=28)} = 2.05$ is found using TableA.2 in the text. Press [+] 2.05 [*] [$\sqrt{}$] [(] 49.76 [$x^2$] [+] 15.6391 [$x^2$] [)] and press [ENTER] to evaluate the upper limit of the prediction interval.

Press [2nd] [ENTRY] to place the expression on the homescreen. Use the arrow keys, [▲] [◄] to place the cursor on the [+] sign to the right of Y1. Press [-] to change the [+] to a [-] sign. Press [ENTER] to obtain the lower limit of the prediction interval. The results are shown in Figure 14.27.

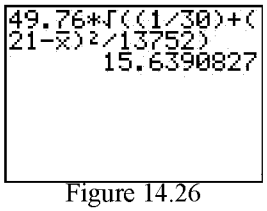


Figure 14.26

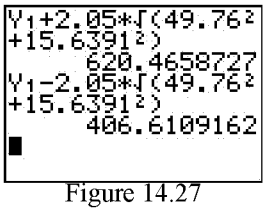


Figure 14.27

Figure 14.27 indicates the probability is 0.95 that a randomly selected 21-year-old will read the sign at somewhere between roughly 407 and 620 feet.

There will be variation among 21-year-olds with regard to the sign-reading distance, so the predicted distance of 514 feet is not likely to be the exact distance for the next 21-year-old who views the sign. Rather than predicting that the distance will be exactly 514 feet, we should instead predict that the distance will be within a particular interval of values. A 95% prediction interval describes the values of the response variable ($y$) for 95% of all individuals with a particular value of $x$. This interval can be interpreted in two equivalent ways:
1. The 95% prediction interval estimates the central 95% of the values of $y$ for members of the population with a specified value of $x$.
2. The probability is 0.95 that a randomly selected individual from the population
with a specified value of $x$ falls into the corresponding 95% prediction interval.

## Estimating the Mean $y$ at a Specified $x$

Thus far in this section we have focused on the estimation of the values of the response variable for individuals. A researcher may instead want to estimate the mean value of the response variable for individuals with a particular value of the explanatory variable. We might ask, "What is the mean weight for college men who are 6 feet tall?" This question only asks about the mean weight in a group with a common height, and it is not concerned with the deviations of individuals from that mean.
In technical terms, we wish to estimate the population mean $E(Y)$ for a specific value of $x$ that is of interest to us. To make this estimate, we use a confidence interval.The format for this confidence interval is

$$\text{Sample statistic} \pm \text{Multiplier} \times \text{Standard error}$$

The sample statistic of $E(Y)$ is the value of $\widehat{y}$ determined by substituting the $x$ value of interest into $\widehat{y} = b_o + b_1 x$, the least-squares regression line for the sample. The *standard error* of $\widehat{y}$ is the $s.e.(fit)$ shown in the Tech Note box in the previous section of the text. The formula for the confidence interval for $y$ at a specific $x$ is

$$\widehat{y} \pm t^* s.e.(fit)$$

where

$$s.e.(fit) = s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})2}}$$

The "multiplier" $t^*$ is found using a t-distribution with $df = n-2$. TableA.2 can be used to find the multiplier for the desired level for the interval.

**Example 14.5 - Continued - Driver Age and Highway Sign-Reading Distance**

Previously in Example 14.5, you tested a hypotheses about the population slope for the regression of sign-reading distance and driver age using your TI calculator. You placed the function Y1 in REgEq: in the linear regression t-test screen. That function Y1 now stores the sample regression equation: $\widehat{y} = 577 - 3.01x$.

**Follow these steps** to obtain a confidence interval for the mean maximum distance at which an individual can read a highway sign by substituting his or her age for x in the sample regression equation.

1. Place a value of x in the sample regression equation.

   Press [2][1][STO→][x,T,θ,n][ENTER] to store 21 in the x register, as shown in Figure 14.28.

2. Evaluate sample regression equation: $\widehat{y} = 577 - 3.01x$, where x is replaced with 21.

   Press [VARS][▶][1][1][ENTER] to place the function Y1 on the homescreen. Press [ENTER] to evaluate Y1. The results are shown in Figure 14.29.

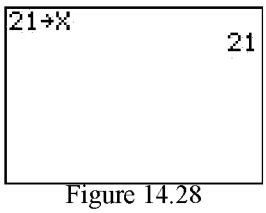


Figure 14.28

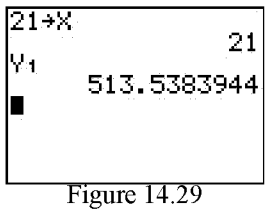


Figure 14.29

For a person 21 years old, the predicted distance is approximately $\widehat{y}$ = 577 - 3(21) = 514 feet.

**Follow these steps** to obtain a 95% confidence interval of the mean maximum distance at which a 21 year-old individual can read a highway sign.

1. Find $s.e.(fit) = s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})2}}$. Recall $s = \sqrt{\frac{SSE}{n-2}} = 49.76$ and $\sum(x_i - x)^2 = 13752$ were found previously. In addition $n = 30$.

   Press 49.76[*][2nd][√][(][(][1][÷][30][)][+][(]21[VARS][5][2][)][$x^2$][÷]13752[)] and press [ENTER] to execute the command.The expression and results are shown in Figure 14.30. The $s.e.(fit) = 15.6391$.

2. Obtain the 95% confidence interval of the mean maximum distance at which a 21 year-old individual can read a highway sign.

Press [VARS] [►] [1] [1] [ENTER] to place the function Y1 on the homescreen. Recall that the "multiplier" $t^*_{(\frac{\alpha}{2}=0.05, df=28)} = 2.05$ is found using TableA.2 in the text. Press [+] 2.05 [*] 15.6391 and press [ENTER] to evaluate the upper limit of the confidence interval.

Press [2nd] [ENTRY] to place the expression on the homescreen. Use the arrow keys, [▲] [◄] to place the cursor on the [+] sign to the right of Y1. Press [-] to change the [+] to a [-] sign. Press [ENTER] to obtain the lower limit of the 95% confidence interval. The results are shown in Figure 14.27.

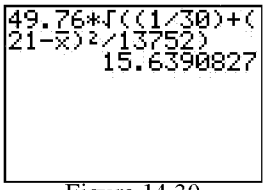



Figure 14.30

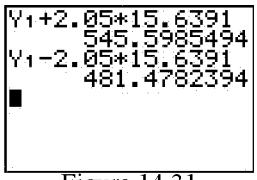



Figure 14.31

Figure 14.31 indicates the probability is 0.95 that the mean distance that a 21-year-old individual will read the sign at somewhere between roughly 481.5 and 545.6 feet.

It is important to note that in.regression the difference between a prediction interval for $y$ and a confidence interval for $E(y)$ is as follows.
A prediction interval for $y$ predicts the value of the response variable $(y)$ for an individual with a particular value of $x$. This interval also estimates the range of values for a specified central percentage of a population of individuals with the same particular value of $x$.
A confidence interval for $E(y)$ estimates the mean value of $y$ for a population of individuals who all have the same particular value of $x$.

## 14.5 Checking Conditions For Regression Models

There are a few conditions that should be at least approximately true when we use a regression model to make an inference about a population. Of the five conditions that follow,the first two are particularly crucial.

**Conditions for Linear Regression**

1. The form of the equation that links the mean value of y to x must be correct. For instance, we won't make proper inferences if we use a straight line to describe a curved relationship.
2. There should not be any extreme outliers that influence the results unduly.
3. The standard deviation of the values ofy from the mean y is the same regardless of the value of the x variable. In other words, y values are similarly spread out at all values of x.
4. For individuals in the population with the same particular value of x, the

distribution of the values of y is a normal distribution. Equivalently, the distribution of deviations from the mean value of y is a normal distribution. This condition can be relaxed if the sample size is large.
5. Observations in the sample are independent of each other.

## Checking the Conditions with Plots

A scatterplot of the raw data and plots of the residuals provide information about the validity of the assumptions. Remember that a residual is the difference between an observed value and the predicted value for that observation and that some assumptions made for a linear regression model have to do with how $y$ values deviate from the regression line. If the properties of the residuals for the sample appear to be consistent with the mathematical assumptions made about deviations within the population, we can use the model to make statistical inferences.

**Example 14.3 - Continued - Relationship Between Height and Weight for College Men**

In a statistics class, the height and weight of a sample of men were recorded. The data is displayed in Table 14.2

| Student | Height | Weight | Student | Height | Weight |
|---|---|---|---|---|---|
| 1 | 73 | 195 | 23 | 72 | 230 |
| 2 | 69 | 135 | 24 | 74 | 170 |
| 3 | 70 | 145 | 25 | 68 | 151 |
| 4 | 72 | 170 | 26 | 73 | 220 |
| 5 | 73 | 172 | 27 | 68 | 145 |
| 6 | 69 | 168 | 28 | 70 | 130 |
| 7 | 68 | 155 | 29 | 72 | 160 |
| 8 | 71 | 188 | 30 | 70 | 210 |
| 9 | 71 | 175 | 31 | 67 | 145 |
| 10 | 68 | 158 | 32 | 67 | 185 |
| 11 | 69 | 185 | 33 | 71 | 237 |
| 12 | 67 | 146 | 34 | 72 | 205 |
| 13 | 66 | 135 | 35 | 73 | 147 |
| 14 | 67 | 150 | 36 | 68 | 170 |
| 15 | 72 | 160 | 37 | 72 | 181 |
| 16 | 68 | 155 | 38 | 68 | 150 |
| 17 | 75 | 230 | 39 | 67 | 150 |
| 18 | 68 | 149 | 40 | 70 | 200 |
| 19 | 73 | 240 | 41 | 71 | 175 |
| 20 | 72 | 170 | 42 | 70 | 155 |
| 21 | 72 | 198 | 43 | 67 | 167 |
| 22 | 72 | 163 | | | |

Table 14.2

**Follow these steps** to obtain the regression equation and to examine the residual plots.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

   c. Turn Diagnostics On to display $r$, the correlation coefficient, and $r^2$, the coefficient of determination.

   Press [2nd] [CATALOG], located on the bottom row, 2nd column from the left above the [0]. Press [ALPHA] [D], and use the down arrow key to locate **DiagnosticOn**. Press [ENTER] to select the command and press [ENTER] once again to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the "height" data: 73, 69, 70, ... pressing [ENTER] after each entry.
   Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the "weight" data: 195, 135, 145, ... in L2 pressing [ENTER] after each entry, as shown in Figure 14.1.

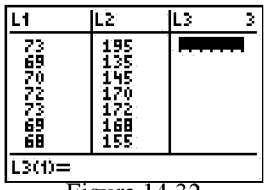



Figure 14.32

3. Obtain the regression equation.

   Press [STAT] [▶] to obtain the [STAT] CALC menu.

   a. Use the down arrow key, [▼], seven times and press [ENTER], or just press [8] to select 8: LinReg (a+bx), as shown in Figure 14.33. Press [2nd] [L1] to

select the "height" data. Press , 2nd L2 to select the "weight" data, as shown in Figure 14.34. Press ENTER to execute the command. The output from the TI calculator is displayed in Figure 14.4.

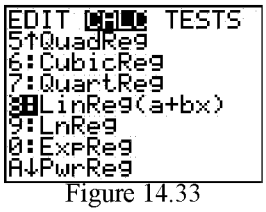


Figure 14.33

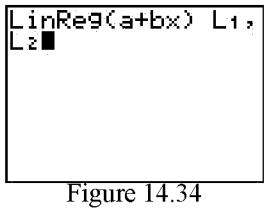


Figure 14.34

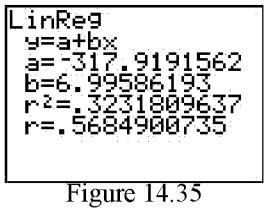


Figure 14.35

The sample regression equation is $\widehat{y} = -317.9 + 6.996x$; the $y$-intercept is $-317.9$ and the slope is $+6.996$. The correlation coefficient, $r = 0.568$ describes a moderately weak positive association. The squared correlation is $r = (0.568)^2 = 0.323$, indicating that there is considerable variation from the line at any given height.

4. Create a plot of residuals versus x and a histogram of residuals.

Press 2nd STAT PLOT accessing the stat plot menu.
Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the first row, the scatterplot. Press ENTER. Use the down arrow key to select list L1 as the list, 2nd L1. Use the down arrow key to enter list RESID as the Ylist: 2nd LIST, using the down arrow key, ▼, until the RESID list is located. Press ENTER to select the list of residuals, RESID. Use the down arrow key to select the first icon for the mark. The settings for Plot 1 are shown in Figure 14.36. When you execute a regression model, the automatic residual list calculates and stores the residuals to the list name RESID. RESID is then an item on the LIST NAMES menu.

5. View the graph.

Press ZOOM, 9: ZoomStat to view the graph, as shown in Figure 14.37.

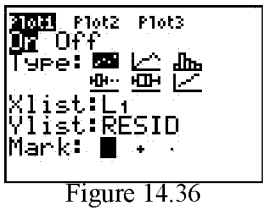


Figure 14.36

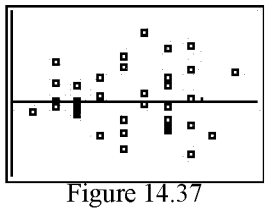


Figure 14.37

In Figure 14.37 there is a plot of the residuals ($e_i$) versus the corresponding values of height for these 43 men. This plot is further evidence that the right model has been used. If the right model has been used, the way in which individuals deviate from the line (residuals) will not be affected by the value of the explanatory variable. The somewhat random-looking blob of points in Figure 14.37 is the way a plot of residuals versus $x$ should look if the right equation

for the mean has been used.

6. Create a histogram of residuals.

   Press [2nd] [STAT PLOT] accessing the stat plot menu.
   Press [ENTER], selecting Plot 1. Place the cursor on ON and press [ENTER]. Use the down arrow key and the right arrow key to select the third icon in the first row, the histogram. Press [ENTER]. Use the down arrow key to enter list RESID as the Xlist: [2nd] [LIST], using the down arrow key, [▼], until the RESID list is located. Press [ENTER] to select the list of residuals, RESID. Use the down arrow key, [▼], entering 1 for the Freq:. Use the down arrow key to select the first icon for the mark. The settings for Plot 1 are shown in Figure 14.38.

7. Set the Window viewing variables in order to view the graph.

   Press [WINDOW], row 1, column 2. Set Xmin to -55. Set Xmax to 70; Xscl to 10; Ymin to -1. Set Ymax to 13; Yscl to 1; Xres to 1. These settings are illustrated in Figure 14.39

8. View the graph.

   Press [GRAPH] to view the graph, as shown in Figure 14.40.

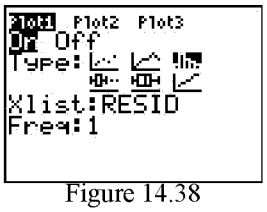


Figure 14.38

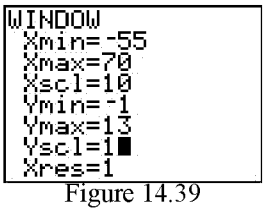


Figure 14.39

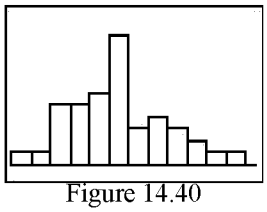


Figure 14.40

To examine the distribution of the deviations from the line, a histogram of the residuals is useful, although for small samples a histogram may not be informative. A more advanced plot called a normal probability plot can also be used to check whether the residuals are normally distributed, but we do not provide the details in this text. Figure 14.40 displays a histogram of the residuals for Example 14.3.It appears that the residuals are approximately normally distributed.

## Corrections When Conditions are Not Met

If attempts are made to use a straight line to describe a curved relationship, a transformation may be required. This is equivalent to using a different model. Fortunately, often the same transformation will correct problems with Conditions 1, 3, and 4. For instance when the response variable is monetary, such as salaries, it is

often more appropriate to use the relationship

$$\ln(y) = b_0 + b_1 x + \epsilon$$

In other words, to assume that there is a linear relationship between the natural log of y and the x values. This is call a log transformation. The following example uses log transformations on both y and x values.

**Example - Not in the Text - Log Transformations**

In a ecology class, students recorded the presence of contaminants in a variety of vegetables from gardens bordering well-traveled highways. The level of lead is one of the contaminants. The data is displayed in Table 14.3

| Distance | Contaminant |
|---|---|
| 5 | 710 |
| 10 | 660 |
| 50 | 265 |
| 100 | 125 |
| 200 | 86 |
| 500 | 65 |
| 750 | 38 |

Table 14.3

**Follow these steps** to perform log transformations and obtain the regression equation.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.
   c. Turn Diagnostics On to display $r$, the correlation coefficient, and $r^2$, the coefficient of determination.

   Press [2nd] [CATALOG], located on the bottom row, 2nd column from the left above the [0]. Press [ALPHA] [D], and use the down arrow key to locate **DiagnosticOn**. Press [ENTER] to select the command and press [ENTER] once again to execute the command.

2. Enter data using the [STAT] list editor.

Press STAT ENTER to select the STAT list editor.

a. Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the "distance" data: 5, 10, 50, ... pressing ENTER after each entry. Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the "contaminant" data: 710, 660, 265, ... in L2 pressing ENTER after each entry, as shown in Figure 14.41.

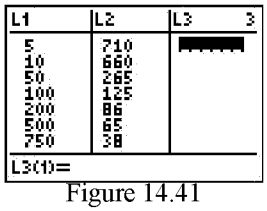


Figure 14.41

3. Plot the statistical data by creating a scatterplot of the distance and contaminant data.

Press 2nd STAT PLOT accessing the stat plot menu.
Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the first row, the scatterplot. Press ENTER. Use the down arrow key to select list L1 as the list, 2nd L1. Use the down arrow key to enter list L2 as the Ylist: 2nd L2. Use the down arrow key to select the second icon for the mark. The settings for Plot 1 are shown in Figure 14.42.

4. View the graph.

Press ZOOM, 9: ZoomStat to view the graph, as shown in Figure 14.43.

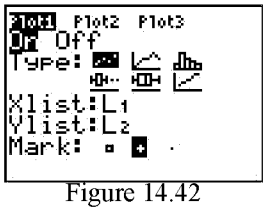


Figure 14.42

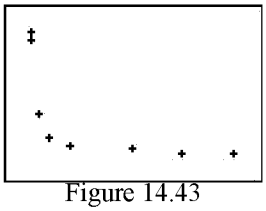


Figure 14.43

The scatterplot in Figure 14.43 clearly indicates that there is not a linear relationship between "distance" and "contaminant."

5. Perform logarithmic transformations of the distance and contaminant data.

Press STAT ENTER to select the STAT list editor.

a. Place the cursor on list L3 at the top of the list. Press LN 2nd L1 ENTER to use a logarithmic transformation of the "distance" data, as shown in Figure 14.44. Figure 14.44 displays the equation before pressing the ENTER key. LN is located on the eighth row, left column.

b. Place the cursor on list L4 at the top of the list. Press LN 2nd L2 ENTER to use a logarithmic transformation of the "contaminant" data, as shown in Figure 14.45. LN is located on the eighth row, left column.

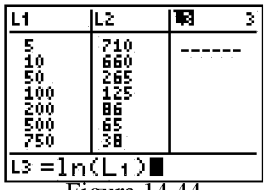


Figure 14.44

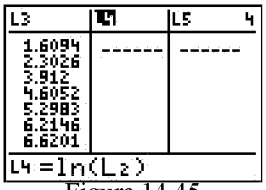


Figure 14.45

6. Obtain the regression equation.

Press STAT ► to obtain the STAT CALC menu.

Use the down arrow key, ▼, seven times and press ENTER, or just press 8 to select 8: LinReg (a+bx), as shown in Figure 14.46. Press 2nd L3 to select the natura; logarithm of the"distance" data. Press , 2nd L4 to select the natural logarithm of the "contaminant" data, as shown in Figure 14.47. Press ENTER to execute the command. The output from the TI calculator is displayed in Figure 14.4.

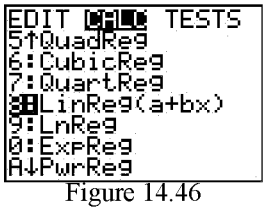


Figure 14.46

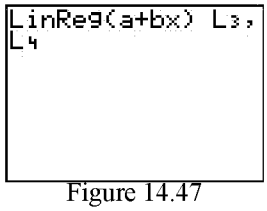


Figure 14.47

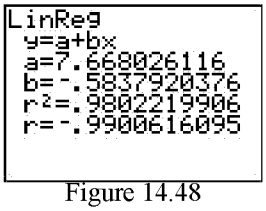


Figure 14.48

The sample regression equation is $\widehat{y} = 7.668 - 0.584x$; the $y$-intercept is $7.668$ and the slope is $-0.584$. The correlation coefficient, $r = 0.991$ describes a very strong positive association. The squared correlation is $r = (0.991)^2 = 0.980$, indicating that there is almost no variation from the line at any given "distance".

7. Obtain the predicted values from the regression equation.

Press STAT ENTER to select the STAT list editor.

Place the cursor on list L5 at the top of the list.
Press 2nd $e^x$ 7.668-0.584* 2nd L3 ) ENTER to reverse our logarithmic transformations. Figure 14.49 displays the equation before pressing the ENTER

key. $\boxed{e^x}$ is located on the eighth row, left column, above $\boxed{\text{LN}}$.

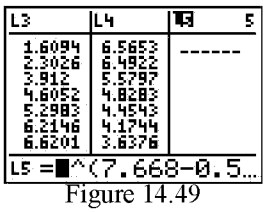



Figure 14.49

8. Create a scatterplot of the "distance" and "contaminant" data and a xyLine graph of the "distance" versus the predicted values of the "contaminant". data.

   Press $\boxed{\text{2nd}}\boxed{\text{STAT PLOT}}$ accessing the stat plot menu.
   Press $\boxed{\text{ENTER}}$, selecting Plot 1. Place the cursor on ON and press $\boxed{\text{ENTER}}$. Use the down arrow key and the right arrow key to select the first icon in the first row, the scatterplot. Press $\boxed{\text{ENTER}}$. Use the down arrow key to select list L1 as the Xlist, $\boxed{\text{2nd}}\boxed{\text{L1}}$. Use the down arrow key to enter list L2 as the Ylist: $\boxed{\text{2nd}}\boxed{\text{L2}}$. Use the down arrow key to select the second icon for the mark. The settings for Plot 1 are shown in Figure 14.50.

   Use the up arrow and the right arrow keys to select Plot 2. Place the cursor on ON and press $\boxed{\text{ENTER}}$. Use the down arrow key and the right arrow key to select the second icon in the first row, the xyLine plot. Press $\boxed{\text{ENTER}}$. Use the down arrow key to select list L1 as the Xlist, $\boxed{\text{2nd}}\boxed{\text{L1}}$. Use the down arrow key to enter list L2 as the Ylist: $\boxed{\text{2nd}}\boxed{\text{L2}}$. Use the down arrow key to select the second icon for the mark. The settings for Plot 2 are shown in Figure 14.51.

9. View the graph.

   Press $\boxed{\text{ZOOM}}$, 9: ZoomStat to view the graph, as shown in Figure 14.52.

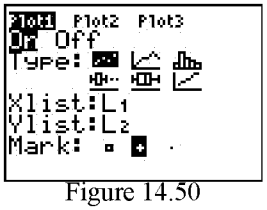



Figure 14.50

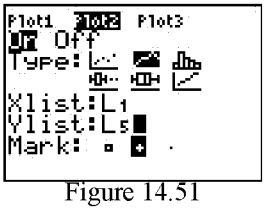



Figure 14.51

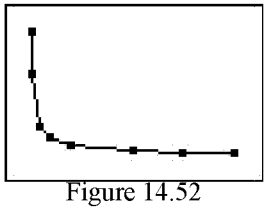


Figure 14.52

The scatterplot in Figure 14.52 clearly indicates that there is an excellent fit between "distance" versus "contaminant" data and the "distance" versus the predicted values of the "contaminant". data. In Figure 14.48, the TI output indicates the sample regression equation is $\widehat{y} = 7.668 - 0.584x$. If we go back to the original data and apply the transfomation the sample regression equation is $\widehat{y} = e^{(7.668 - 0.584 * \ln(distance))}$. The TI output in Figure 14.52 indicates that we have been successful in modelling this data.

# Chapter 15
# More About Inference For Categorical Variables

In Chapter 6, we learned techniques for analyzing a fundamental question that researchers often ask about two categorical variables, which is

*Is there a relationship between the variables so that the chance that an individual falls into a particular category for one variable depends upon the particular category they fall into for the other variable?*

For instance, in Example 6.1 we saw that the likelihood of divorce is greater for smokers than it is for nonsmokers. The chi-square test is a procedure for assessing the statistical significance of a relationship between categorical variables. We introduced the basic ideas of this statistical method in Chapter 6, and in this chapter we examine the chi-square test more fully. We will use the chi-square procedure to test hypotheses about two-way tables, and also we will learn how the chi-square procedure is used to test hypotheses about a single categorical variable. .

After reading this chapter you should be able to:

1. Conduct a chi-square test, including observed counts, computing a chi-squared statistic, and finding the $p$-value.
2. Find a $p$-value given the chi-square value and degrees of freedom.
3. Use Fisher's Exact Test.for $2 \times 2$ tables.
4. Conduct a chi-square.goodness of fit test, finding expect counts, calculating the test statistic and finding the $p$-value.

**Keystrokes Introduced**

1. [2nd] [MATRIX] [►] [►], accessing the EDIT menu. You will enter values into a matrix, laying the foundation for conducting a chi-square test based upon a contingency table.
2. [STAT] [►] [►], accessing the TESTS menu. Select C: $\chi^2$-Test. This test will conduct the actual chi-square test based upon a contingency table.
3. [2nd] [MATRIX] >NAMES, selecting 2: $[B]$. You will use this command to examine the expected counts in a chi-squared test.
4. [MATH] [►] [►] [►], accessing the PRB menu, selecting 3: $nCr$. You will use this command to use Fisher's Exact Test.for $2 \times 2$ tables.
5. [2nd] [TBLSET] will be used to set values in a table.
6. [2nd] [TABLE] will be used to examine a table.

7. VARS ▶, accessing the Y-VARS menu, selecting 1: Function. You will use this command to find a probability.

## 15.1 The Chi-Square Test for Two-Way Tables

In Chapter 6,we learned these terms:

1. The raw data from a categorical variable consist of group or category names that don't necessarily have any logical ordering. Your sex (male or female) and your handedness (left or right), for instance, are both categorical variables.
2. A two-way table displays the counts of how many individuals fall into each possible combination of categories of two categorical variables. A two-way table may also be called a contingency table.
3. Each combination of a row and a column is referred to as a **ce1l** of the contingency table.
4. **Row percents** are the percents across a row of a contingency table. They are based on the total number of observations in the row.
5. **Column percents** are the percents down a column of a contingency table. They are based on the total number of observations in the column.

When we consider two categorical variables, a question of interest is whether a relationship observed in a sample can be used to infer that there is a relationship in the population(s) from which the sample(s) was drawn. This question can be answered using a procedure called a chi-square test.

**Old Example 15.1 - Ear Infections and Xylitol Sweetner**

Xylitol is a food sweetner that may also have antibacterial properties. In an experiment conducted in Finland, researchers investigated whether the regular use of chewing gum containing Xylitol could reduce the risk of a middle ear infection for children in daycare centers (Uhari et al., 1998). The investigators randomly divided 533 children in daycare centers into three groups. One group regularly chewed gum that contained Xylitol, another group regularly took Xylitol lozenges, and the third group regularly chewed gum that did not contain Xylitol. The experiment lasted for three months and, for each child, the researchers recorded whether the child had an ear infection during that period. The data is shown in Table 15.1:

| | Ear Infection in Three Months | |
|---|---|---|
| Group | No | Yes |
| Placebo Gum | 129 | 49 |
| Xylitol Gum | 150 | 29 |
| Xylitol Lozenge | 137 | 39 |

Table 15.1

The purpose of the experiment is to compare the three groups with regard to the

proportion of children who experienced an ear infection. The goal is to see if the sample results allow us to infer that differences would exist in the population of all children if they were to be administered these treatments. For this purpose there are three relevant parameters.

$p_1$ = proportion who would get an ear infection in a population given the placebo gum
$p_2$ = proportion who would get an ear infection in a population given Xylitol gum
$p_3$ = proportion who would get an ear infection in a population given Xylitol lozenges

If the chance of falling into the ear infection category is not related to treatment method, the risk of an ear infection would be the same for the three treatments. So, null and alternative hypotheses are

$H_0 : p_1 = p_2 = p_3$ or (No relationship between treatment and outcome.)
$H_a : p_1, p_2, p_3$ are not all the same. (There is a relationship.)

Notice that the alternative hypothesis simply states that the three proportions are not all the same. A limitation of chi-square tests is that no particular direction for the relationship can be stated in the alternative hypothesis.

**Follow these steps** to conduct a chi-square test, including observed counts, computing a shi-square statistic and finding the $p$-value.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Turn off all StatPlots.

   Press [2nd] [STAT PLOT] accessing the StatPlot menu. Select 4: PlotsOff, pressing [ENTER] and [ENTER] again to turn all plots off.
   c. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter the observed counts using the [MATRIX] list editor.

   Press [2nd] [MATRIX] [▶] [▶], accessing the EDIT menu, as shown in Figure 15.1. Select 1: $[A]$, as shown in Figure 15.1

   a. Enter the dimensions of the matrix. This matrix has 3 rows and 2 columns. Press [2] [ENTER] [3] [ENTER]. Observe that the cursor is now on the first

row first column as shown in Figure 15.2.

b. Enter the observed counts.

Type the element in the first row, first column: 129 and press ENTER.
Type the element in the first row, second column: 49 and press ENTER.
Type the element in the second row, first column: 150 and press ENTER.
Type the element in the second row, second column: 29 and press ENTER.
Type the element in the third row, first column: 137 and press ENTER.
Type the element in the third row, second column: 39 and press ENTER.
The resulting matrix is shown in Figure 15.3

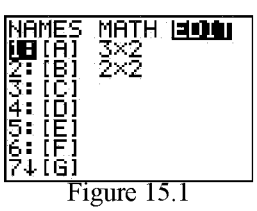


Figure 15.1

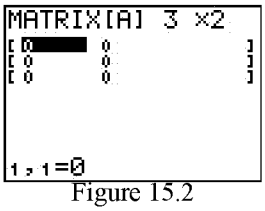


Figure 15.2

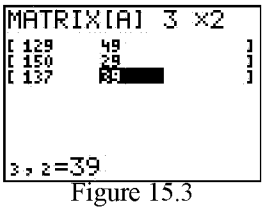


Figure 15.3

3. Compute the chi-square statistic.

Select STAT ► ►, accessing the TESTS menu. Select C: $\chi^2$-Test, as shown in Figure 15.4, and press ENTER. Place matrix A in Observed: by pressing MATRIX >NAMES, selecting 1: $[A]$, as shown in Figure 15.5. Press ENTER. Place matrix B in Expected: by pressing MATRIX >NAMES, selecting 2: $[B]$, as shown in Figure 15.5. Press ENTER. Use the down arrow key, highlighting Calculate and ENTER to execute the command. The results are shown in Figure 15.6.

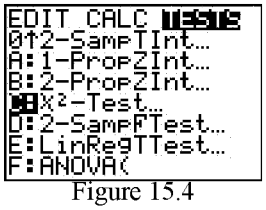


Figure 15.4

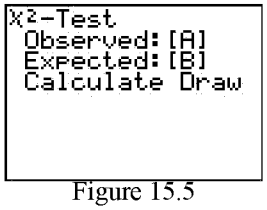


Figure 15.5

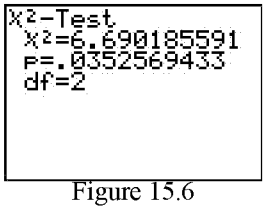


Figure 15.6

The TI output, as shown in Figure 15.6 shows that the p-value is 0.035. As a result, we can conclude that there is a statistically significant relationship between the risk of an ear infection and the preventative treatment used.

4. Examine the expected counts.

Press 2nd MATRIX >NAMES, selecting 2: $[B]$, as shown in Figure 15.7. Press ENTER ENTER to display the expected counts. Use the right arrow key, ►, to examine the expected counts for each row and column, as shown

in Figure 15.8.

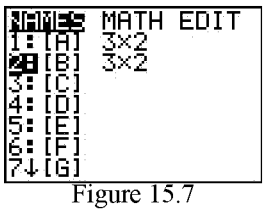
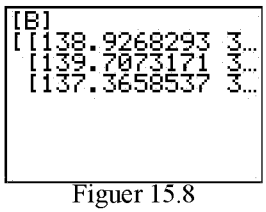


Figure 15.7



Figuer 15.8

5. Verify the $p$-value.

On the homescreen, press 2nd DISTR, located on the fourth row, column four, above VARS. Select 7: $\chi^2$cdf( , as shown in Figure 15.9.
Type 6.6902,1E99 , 2 ) ENTER. The arguments in the normalcdf function are *lowerbound,upperbound,df*. The lowerbound is 6.6902; the upperbound is $1E99$, or $1 \times 10^{99}$, giving a very large value of $\chi^2$ to the right of the observed $\chi^2$-statistic. The degrees of freedom is 2. Press ENTER. The results are shown in Figure 15.10, indicating the $p$-value is 0.0353, or 0.035 rounded to 3 decimal places.

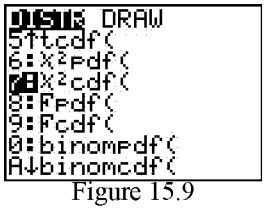


Figure 15.9

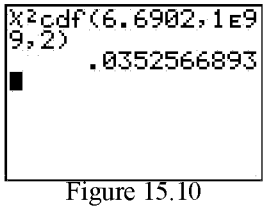


Figure 15.10

## 15.2 Fisher's Exact Test for 2×2 Tables

For 2 x 2 tables, Fisher's Exact Test is another test that can be used to another test that can be used to analyze the statistical significance of the relationship. In theory, this test can be used for any 2 x 2 table, but most commonly it is used when necessary sample size conditions for using the z-test or the chi-square test are violated. In large part, this has to do with computational difficulties that occur for large sample sizes.
Although the computations are cumbersome, the idea behind Fisher's Exact Test is easy. Suppose, for example, that a randomized experiment is done to see if taking the herb echinacea reduces the risk of getting a cold, and the observed data are 1of 10 people taking echinacea get a cold during the study while 4 of 10 taking a placebo get a cold. In all, 5 of the 20 participants got a cold, but only 1 was in the echinacea group. The p-value for a one-sided Fisher's Exact Test is the answer to this question:

*Given that 5 of 20 participants get a cold regardless of treatment method, what is the probability that just 1 or fewer would be in the echinacea group?*

Notice that the test statistic is simply the count of how many got a cold in the echi-

nacea group. As always, the $p$-value question addresses how likely it is that the test statistic would be as extreme or more extreme as it is in the direction of the alternative hypothesis if the null hypothesis is true. The answer (it's 0.152) is determined using a probability distribution called the hypergeometric distribution. Values of the hypergeometric probability density function are found in the following manner:

> A population consists of $N$ units with $a$ of Type A and $b$ of Type B. A random sample on $n$ units is selected, and $X$ = number of Type A in the sample. Then $X$ has a hypergeometric probability distribution and the probability distribution is
>
> $$P(X = k) = \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{N}{n}}$$
>
> for k = 0, 1, ..., smaller of (a, n)

**Follow these steps** to conduct Fisher's Exact Test.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].
   b. Turn off all StatPlots.

   Press [2nd] [STAT PLOT] accessing the StatPlot menu. Select 4: PlotsOff, pressing [ENTER] and [ENTER] again to turn all plots off.
   c. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Use the hypergeometric probability density function to answer this question:

   *Given that 5 of 20 participants get a cold regardless of treatment method, what is the probability that just 1 or fewer would be in the echinacea group?*
   a. Press [Y=].
   b. There are 10 people taking echinacea and 10 taking a placebo, for a total of 20 people in the study. In all, 5 of the 20 participants got a cold. So, we enter into \Y1= :
   [(] [1] [0] [MATH] [►] [►] [►], accessing the PRB menu, as shown in Figure 15.11. Select 3: $nCr$, and press [$X$,$T$,$\theta$,$n$] [)] [×] as shown in Figure 15.12 Continue with [(] [1] [0] [MATH] [►] [►] [►], accessing the PRB menu, as

shown in Figure 15.11. Select 3: $nCr$, and press [5] [−] [$X$,$T$,$\theta$,$n$] [)] [)] [÷] as shown in Figure 15.13

Continue with [(] [2] [0] [MATH] [►] [►] [►], accessing the PRB menu, as shown in Figure 15.11. Select 3: $nCr$, and press [5] [−] [$X$,$T$,$\theta$,$n$] [)] [)] [÷] as shown in Figure 15.14

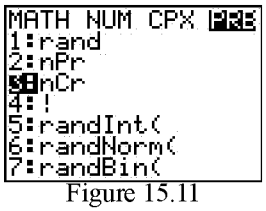


Figure 15.11

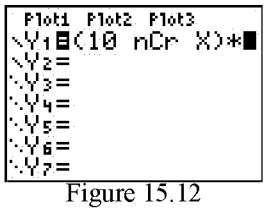


Figure 15.12

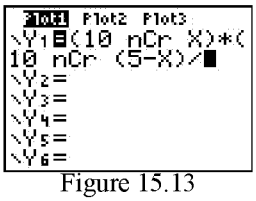


Figure 15.13

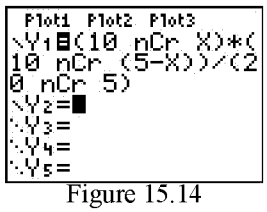


Figure 15.14

c. Set the values in the Table.

Press [2nd] [TBLSET], located above the WINDOW key. Let TblStart = 0. Use the down arrow key, [▼], letting ΔTbl=1. Use the down arrow key, [▼], letting Indpnt: be AUTO. Use the down arrow key, [▼], letting Depend: be AUTO, as shown in Figure 15.15

d. Examine values in the table:

Press [2nd] [TABLE], located above the GRAPH key. Identiy that $P(x = 0) = 0.1625$ and $P(x = 1) = 0.13545$, as shown in Figure 15.16.

Figure 15.15

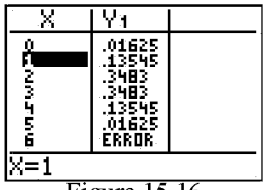


Figure 15.16

e. Find the probability that *the probability that just 1 or fewer would be in the echinacea group* by adding $P(x = 0)$ and $P(x = 1)$.

Press [VARS] [►], accessing the Y-VARS menu. Select 1: Function and press [ENTER]. Select 1: $Y_1$ and press [ENTER]. Press [(] [0] [)] [+].

Continue with

Pressing [VARS] [►], accessing the Y-VARS menu. Select 1: Function and press [ENTER]. Select 1: $Y_1$ and press [ENTER]. Press [(] [1] [)] Press

ENTER to execute the command. The results are shown in Figure 15.17.

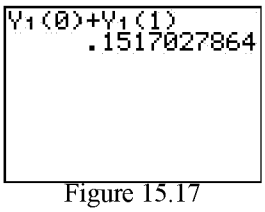



Figure 15.17

The p-value for a one-sided Fisher's Exact Test is the answer to this question:

*Given that 5 of 20 participants get a cold regardless of treatment method, what is the probability that just 1 or fewer would be in the echinacea group?*

As always, the $p$-value question addresses how likely it is that the test statistic would be as extreme or more extreme as it is in the direction of the alternative hypothesis if the null hypothesis is true. The answer is 0.152.

## 15.3 One Categorical Variable: $\chi^2$ Goodness of Fit

We sometimes wish to test whether the probabilities of falling into the possible categories of a single categorical variable are given by a specified set of values. For example, if a lottery drawing involves randomly drawing digits between 0 and 9, we may wish to analyze observed data to confirm that the probability over the long run is 1/10 for drawing each of the ten digits from 0 to 9. With two modifications from how it is applied to two-way tables, the chi-square statistic can be used to test hypotheses about the probability distribution of a single categorical variable. In such instances, the significance test is called the chi-square goodness of fit test.

**Old Example 15.13 - The Pennsylvania Daily Number**

The Pennsylvania Daily Number is a state lottery game in which the state constructs a three-digit number by drawing a digit between 0 and 9 from each of three different containers. If the digits drawn, in order, were 3,6, and 3, for example,then the daily number would be 363. In this example,we focus only on draws from the first container. If numbers are randomly selected, each value between 0 and 9 would be equally likely to occur. This leads to the following null hypothesis:

$H_0 : p = \frac{1}{10}$ for each of the 10 possible digits in first container

Simply stated, the alternative hypothesis is that the null hypothesis is false. In this setting, that would mean that the probability of selection for some digits is different from 1/10.

**Follow these steps** to conduct a $\chi^2$ Test of Goodness of Fit.

1. Preparations:

a. Turn off all "Y=" functions.

Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

b. Turn off all StatPlots.

Press [2nd] [STAT PLOT] accessing the StatPlot menu. Select 4: PlotsOff, pressing [ENTER] and [ENTER] again to turn all plots off.

c. Clear all lists in the Stat editor.

Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

2. Enter data using the [STAT] list editor.

Press [STAT] [ENTER] to select the [STAT] list editor.

a. Enter the observations in list L1:

Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the frequencies of the draws: 47, 50, 55,46, 53,39, 55, 55, 44, 56 pressing [ENTER] after each entry, as shown in Figure 15.18.

b. Enter the expected values in list L2
Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the expected values of the draws: enter 50 ten times pressing [ENTER] after each entry, as shown in Figure ??.??.

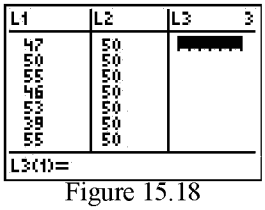


Figure 15.18

3. Calculate the $\chi^2$ test statistic on the homescreen.

Press [2nd] [LIST] [►] [►], accessing the MATH menu. Select 5: sum ( and press [ENTER].

Press a second parenthesis, [(], and [2nd] [L1] [−] [2nd] [L2] [)] [$x^2$] [÷] [2nd] [L2] [)]. Press [ENTER] to execute the command. The results are shown in

Figure 15.19.

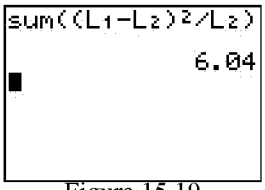



Figure 15.19

4. Find the $p$-value.

   On the homescreen, press 2nd DISTR, located on the fourth row, column four, above VARS. Select 7: $\chi^2$cdf( , as shown in Figure 15.20. Type 6.04,1E99 , 9 ) ENTER. The arguments in the normalcdf function are *lowerbound,upperbound,df*. The lowerbound is $6.04$; the upperbound is $1E99$, or $1 \times 10^{99}$, giving a very large value of $\chi^2$ to the right of the observed $\chi^2$-statistic. The degrees of freedom is $9$. Press ENTER. The results are shown in Figure 15.21, indicating the $p$-value is $0.7359$, or $0.736$ rounded to 3 decimal places.

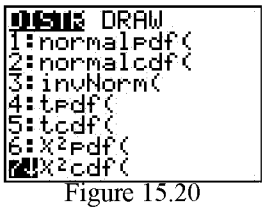



Figure 15.20

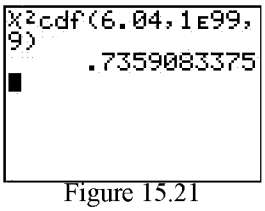



Figure 15.21

The result is not statistically significant. The null hypothesis is not rejected.

# Chapter 16
# Analysis of Variance

Suppose that a researcher wants to compare the mean weight loss for three different weight-loss programs. Or, that another researcher wants to compare the mean testosterone levels of men in seven different occupations. What statistical methods can be used to make the desired comparisons? In the data analysis, the researchers might, at some point,.use confidence intervals and significance tests to compare two means at a time (for example, compare mean testosterone levels of occupations 1 and 2, occupations 1 and 3, and so on). Usually, however, an important first step in the analysis of more than two means is to do a significance test to determine if there are any differences at all among the population means being compared. The significance test for doing this is part of a procedure called the **analysis of variance**, which is also sometimes referred to as **ANOVA**.
In this chapter, we focus on the comparison of the means of more than two populations. When different values or levels of a single categorical explanatory variable (weight-loss programs, for instance) define the populations being compared, the ANOVA procedure is called **oneway analysis of variance**. In general, analysis of variance is a versatile tool for analyzing how the mean value of a quantitative response variable is related to one or more categorical explanatory factors.

After reading this chapter you should be able to:

1. Check assumptions before performing a one-way analysis of variance.
2. Perform a one-way analysis of variance.
3. Find a p-value for an $F$-distribution.

**Keystrokes Introduced**

4. [STAT] [►][►] to obtain the [STAT] TEST menu, selecting F: ANOVA(. The function **ANOVA(***list1,list2* [*,list3,...,list20* ]) performs a one-way analysis of variance for comparing the means of two to 20 populations.
5. [2nd] [DISTR] accesses the distribution menu, slecting 9: Fcdf(. The function **Fcdf(***lowerbound, upperbound, numerator df, denominator df*) computes the F-distribution probability between the *lowerbound* and *upperbound* for the specified *numerator df* and *denominator df*.

## 16.1 Comparing Means With an ANOVA F-Test

When we comnare the means of populations represented by independent samples of a quantitative response variable, a null hypothesis of interest is that all means have the same value. An alternative hypothesis is that the means are not all equal.

Notice that this alternative hypothesis does not require that all means must differ from each other. The alternative would be true, for example, if only one of the means were different from the others. If $k$ = the number of populations, the null and alternative hypotheses for comparing population means can be written as

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_a : \text{The means are not all equal}$$

An **$F$-statistic** that arises from a one-way analysis of variance of the sample data is used to test the hypotheses about the population means, and the significance test is called an $F$-test. The $F$-statistic is sensitive to differences among a set of sample means. The greater the variation among the sample means, the larger the value of the test statistic. The smaller the variation among the observed means, the smaller the value of the test statistic. In this section, we are concerned with the general ideas of the $F$-test.

Conceptually, the $F$-statistic can be viewed as follows:

$$F = \frac{\text{Variation among sample means}}{\text{Natural variation within groups}}$$

## Assumptions and Necessary Conditions for the $F$-Test

In the derivation of the $F$-statistic, the assumptions are similar to those for the pooled two-sample t-procedures described in Chapters 12and 13.The assumptions about the populations and samples representing them are
The samples are independent random samples.
The distribution of the response variable is a normal curve within each population.
The different populations may have different means.
All populations have the same standard deviation, $\sigma$.

**Example 16.1 - Classroom Seat Location and Grade Point Average**

Is it true that the best students sit in the front of a classroom, or is that a false stereotype? in surveys done in two statistics classes at the University of California at Davis, students reported their grade point averages and also answered the question, "Where do you typically sit in a classroom (front, middle, back)?" In all, 384 students gave valid responses to both questions, and among these students 88 said the typicall sit in the front, 218 said they typically sit in the middle, and 78 typically sit in the back. We will use the TI calculator to determine if this data set satisfyies the assumptions and necessary conditions for the $F$-test.
The following data are obtained by combining two worksheets UCDavis1 and UC-Davis2. Missing observations were removed; The "front" and "middle" data were "trimmed" by removing the top 1% and bottom 1% of the data. The "back" data were "trimmed" by removing the bottom 1% of the data.

4.00 2.30 2.60 3.39 2.80 2.40 3.00 2.20 3.60 3.16 2.40
3.28 2.81 2.00 2.60 2.50 3.50 3.25 3.16 2.97 3.00 3.01
3.50 3.43 4.00 2.00 2.50 3.17 2.40 2.29 2.10 3.45 2.18
2.80 3.40 3.70 3.00 3.20 3.00 3.10 2.50 3.05 3.50 2.90
3.25 2.58 2.90 3.40 3.92 3.00 2.84 2.80 2.87 2.95 4.00
3.20 2.50 3.50 3.00 2.90 2.04 3.60 3.40 3.00 3.43 2.10
3.00 2.06 3.00 2.40 2.94 2.50 2.50 2.70 3.20 2.67 2.00
2.46

Table 16.1 GPA's for Back

2.80 2.20 3.20 3.31 3.00 2.50 2.00 1.05 2.10 3.25 3.50
2.52 3.00 2.65 0.55 4.00 3.00 3.50 2.87 3.19 1.98 1.90
2.51 3.20 3.04 3.06 2.80 3.50 3.00 2.40 2.02 2.14 3.00
2.23 2.30 2.58 2.66 3.00 3.00 3.30 1.85 2.75 3.12 3.24
3.00 3.10 2.80 3.37 3.50 2.40 2.00 2.70 3.00 2.50 2.17
2.91 3.22 3.80 3.00 2.87 2.00 3.91 2.50 3.50 3.42 3.30
3.60 2.26 3.18 3.20 3.90 2.00 2.66 3.88 3.15 2.70 2.34
2.81 2.82 3.62 3.10 2.80 3.00 2.70 3.40 2.40 2.76 3.00
3.50 4.00 2.60 2.70 3.43 3.77 3.00 3.25 2.83 3.15 3.65
3.60 3.00 2.25 2.54 2.60 2.93 3.00 2.50 3.00 3.51 2.37
3.50 2.90 3.90 2.65 3.50 3.30 3.40 3.60 3.21 3.85 3.20
2.50 2.90 2.60 2.01 3.87 2.50 2.90 3.60 3.90 3.20 3.30
2.80 3.80 3.00 2.23 3.01 2.50 3.00 2.79 3.80 2.80 2.64
3.65 2.54 2.80 3.50 3.10 2.66 1.91 2.80 2.80 2.70 3.17
2.60 2.58 2.50 2.70 3.50 3.00 2.00 3.34 3.83 2.70 3.90
2.33 3.00 3.10 3.00 3.00 3.25 3.26 2.10 2.77 3.70 3.00
3.88 3.00 3.20 3.57 3.00 2.96 2.70 3.80 2.50 3.87 3.40
3.80 2.04 2.80 2.40 2.90 2.38 3.83 2.90 2.91 3.67 3.00
3.50 3.00 3.70 3.33 4.00 3.00 3.20 3.00 3.20 3.40 3.70
3.40 3.10 3.00 3.79 3.81 2.60 3.00 3.00 3.90

Table 16.2 GPA's for Middle

2.62 3.78 2.34 2.50 3.50 2.60 3.84 3.40 3.20 2.50 3.00
3.30 3.20 3.73 3.00 3.98 3.40 4.00 3.40 3.50 3.10 2.50
3.60 2.83 2.10 3.00 2.00 2.47 3.30 3.00 3.88 2.19 2.74
2.99 3.10 3.00 3.86 4.00 3.00 3.11 3.70 3.50 3.00 3.50
2.62 3.70 3.40 2.80 3.70 4.00 3.22 3.26 3.00 3.71 3.26
2.00 3.88 3.25 3.39 3.40 3.33 3.90 3.80 3.24 3.44 2.44
1.92 1.96 3.00 2.05 4.00 3.89 3.50 2.67 3.00 3.10 3.70
3.54 3.21 3.30 3.80 3.00 3.40 3.50 3.60 3.92 2.90 3.90

Table 16.3 GPA's for Front

**Follow these steps** to check assumptions and perform a one-way analysis of variance.

1. Preparations:
   a. Turn off all"Y=" functions.

   Press [Y=] and press [CLEAR] to remove all functions For each line that is not blank, place the cursor on the function and press [CLEAR] Press [2nd] [QUIT].

   b. Clear all lists in the Stat editor.

   Press [STAT], selecting 4: ClrList. Enter each list name: L1, L2, L3, L4, L5, L6. Press [ENTER] to execute the command.

   c. Turn Diagnostics On to display $r$, the correlation coefficient, and $r^2$, the coefficient of determination.

   Press [2nd] [CATALOG], located on the bottom row, 2nd column from the left above the [0]. Press [ALPHA] [D], and use the down arrow key to locate **DiagnosticOn**. Press [ENTER] to select the command and press [ENTER] once again to execute the command.

2. Enter data using the [STAT] list editor.

   Press [STAT] [ENTER] to select the [STAT] list editor.

   a. Place the cursor on list L1 row 1 to make L1(1) the active list row. Enter the "back" data: 4.00, 2.30, 2.60, ... pressing [ENTER] after each entry.
   Place the cursor on list L2 row 1 to make L2(1) the active list row. Enter the "middle" data: 2.80, 2.20, 3.20, ... in L2 pressing [ENTER] after each entry.
   Place the cursor on list L3 row 1 to make L3(1) the active list row. Enter the "front" data: 2.62, 3.78, 2.34, ... in L3 pressing [ENTER] after each entr, as shown in Figure 16.1.

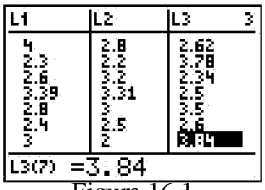


Figure 16.1

3. Create a boxplot of the data to check the assumptions before a one-way analysis of variance is performed.

   Press [2nd] [STAT PLOT] accessing the stat plot menu.

Press ENTER, selecting Plot 1. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select list L1 as the list, 2nd L1. Press ENTER. Use the down arrow key to enter 1 as the Freq:. Press ENTER. Use the down arrow key to select the first icon for the mark. The settings for Plot 1 are shown in Figure 16.2.
Use the up arrow key to select Plot 2. Press ENTER, selecting Plot 2. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select list L2 as the list, 2nd L2. Press ENTER. Use the down arrow key to enter 1 as the Freq:. Press ENTER. Use the down arrow key to select the first icon for the mark. The settings for Plot 2 are shown in Figure 16.3.
Use the up arrow key to select Plot 3. Press ENTER, selecting Plot 3. Place the cursor on ON and press ENTER. Use the down arrow key and the right arrow key to select the first icon in the second row, the modified boxplot. Press ENTER. Use the down arrow key to select list L3 as the list, 2nd L3. Press ENTER. Use the down arrow key to enter 1 as the Freq:. Press ENTER. Use the down arrow key to select the first icon for the mark. The settings for Plot 3 are shown in Figure 16.4.

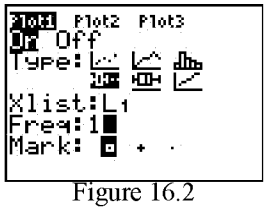


Figure 16.2

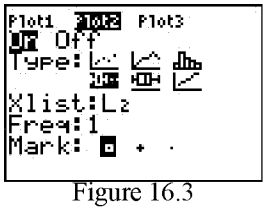


Figure 16.3

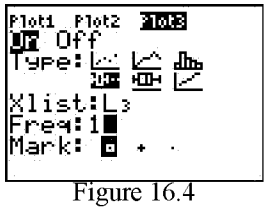


Figure 16.4

4. View the graph.

Press ZOOM, 9: ZoomStat to view the graph, as shown in Figure 14.37.

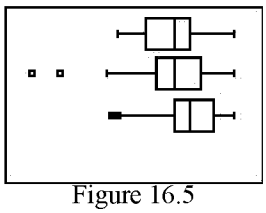


Figure 16.5

In the boxplot, shown in Figure 16.5 we see that students sitting in the front generally have slightly higher GPA's than the others. The boxplot in Figure 16.5 shows two outliers in the group of students who typically sit in the middle of a classroom, but there are 218 students in that group so these outliers don't have much influence on the results. The standard deviations for the three groups are nearly the same, and the data do not appear to be skewed. The necessary conditions for doing an $F$-test are satisfied in this example.

5. Perform a one-way analysis of variance:

   Press STAT ► ► to obtain the STAT TEST menu.

   a. Use the down arrow key, ▼ repeatedly to select F: ANOVA(, as shown in Figure 16.6 and press ENTER. Type 2nd L1 , 2nd L2 , 2nd L3 , to place lists L1, L2, L3 as arguments in the function. These settings are shown in Figure 16.7.Press ENTER to execute the command. The results of the one-way analysis of variance concerning the equality of the population means are shown in Figures 16.8 and 16.9. Use the down arrow key to view the results shown in Figure 16.9.


Figure 16.6


Figure 16.7

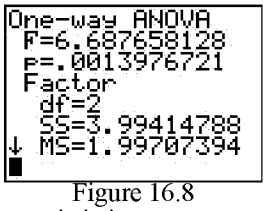



Figure 16.8

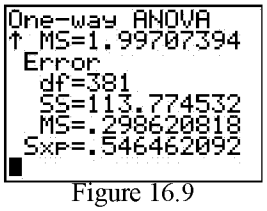



Figure 16.9

The test statistic is

$$F = \frac{\text{Variation among sample means}}{\text{Natural variation within groups}} = \frac{\text{MS}_{\text{Factor}}}{\text{MS}_{\text{Error}}} = \frac{1.997}{0.299} = 6.688$$

The p-value to three decimal places is 0.001.With a p-value this small, we can reject the null hypothesis and thus conclude that there are differences among the means in the populations represented by the samples.

The $df_{Factor} = 2$; the sum of squares between the means, $SS_{Factor} = 3.994$; the $MS_{Factor} = 1.997$. The $df_{Error} = 381$; the sum of squares for error, $SS_{Error} = 113.775$; the $MS_{Error} = 0.299$.

The $df_{Total} = df_{Factor} + df_{Error} = 2 + 381 = 383$. The $SS_{Total} = SS_{Factor} + SS_{Error} = 3.994 + 113.775 = 117.769$.

Note that Tukey's and Fisher's paired comparison methods are not currently available as built in functions on the TI calculators.

## 16.2 The Family of $F$-Distributions

An $F$-distribution is used to find the p-value for an ANOVA $F$-test of the null hypothesis that several population means are equal. The family of F-distributions is a family of skewed distributions, each with a minimum value of 0. A specific $F$ distribution is indicated by two parameters called degreesof freedom The first of

the two parameters is called the numerator degrees of freedom; the second is called the denominator degrees of freedom. The values of the two degrees of freedom parameters always are given in the order numerator $df$, denominator $df$. In one-way ANOVA,the numerator $df = k - 1$ (number of groups - 1), and the denominator $df = N - k$ (total sample size - number of groups).

## Determining the p-Value

In Example 16.1, Figure 16.8, the p-value is reported as part of the output. The TI distribution function $Fcdf$(*lowerbound*, *upperbound*, *numerator df*, *denominator df*) computes the $F$ distribution probability between the *lowerbound* and *upperbound* for the specified *numerator df* and *denominator df*.

**Follow these steps** to obtain the p-value obtained in Example 16.1, Figure 16.8.

1. Find a p-value for an $F$-distribution:

   Press [2nd] [DISTR], using the down arrow key, [▼], several times to select 9: Fcdf(, as shown in Figure 16.10 and press [ENTER]. Type the *lowerbound*, 6.688[,] *upperbound*,1E99, *numerator df*,2, and *denominator df*,381[)]. The *upperbound*,1E99, is translated as $1 \times 10^{99}$. Press [ENTER] to execute the command. The results are shown in Figure 16.11.

Figure 16.10

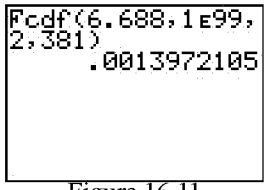


Figure 16.11

   The area to the right of $F = 6.688$ under the F- distribution is the same as the p-value, 0.001.
   Please note that the *df's* must be in the order of *numerator df* and *denominator df*.
   Additional comment:
   Please note that the Kruskal-Wallis test, Mood's median test and the two-way ANOVA are not currently available as built in functions on TI-calculators.

# Troubleshooting the TI-83 and TI-84

If you can't see anything on the screen, you may need to adjust the contrast.
To darken the screen, press, press *and* release [2nd], and then press and hold the up arrow key, [▲], until the display is sufficiently dark.
To lighten the screen, press, press *and* release [2nd], and then press and hold the down arrow key, [▼], until the display is sufficiently light.

## 1 Error Messages

If an error menu is displayed:
Note the error type (ERR: DIM MISMATCH).
Select 2: GOTO, if it is available. The previous screen is displayed with the cursor at or near the error location.
Determine the error and correct the expression if possible.

**Selected Error Conditions**

1. **ARGUMENT** - A function or instruction does not have the correct number of arguments. Consult the TI manual for the function and the correct syntax. For example, if you entered normalcdf(0) and pressed [ENTER], the calculator would return ERR: ARGUMENT. The arguments in [2nd][DISTR] 2: normalcdf( are normalcdf(*lowerbound*,*upperbound*[,$\mu$,$\sigma$]). The correct entry must include the *lowerbound* and *upperbound*: e.g.normalcdf(0,1).

2. **DIM MISMATCH** - Lists or matrices have dimensions that do not match. For example, calculating a regression equation using [STAT][▶], selecting 8: LinReg(a+bx) L1, L2 will produce an ERR: DIM MISMATCH message if the number of elements in L1 and L2 do not match. In order to resolve the error, select [STAT][EDIT], and compare the number of elements in each list L1 and L2. The number of elements must be the same in each of the lists.

3. **STAT PLOT** - You have attempted to display a graph when a stat plot that uses an undefined list is turned on. Press [2nd][STAT PLOT],

selecting 4: PlotsOff, turning off all stat plots. Turn on only the stat plot to be displayed.

4. **SYNTAX** - The command contains a syntax error. Look for misplaced functions, arguments, parentheses, or commas. Consult the TI manual for the function and correct syntax. Select 2: GOTO and recheck your command. Make sure you have pressed $\boxed{X,T,\theta,n}$ for $X$. Make sure that you have distinguished between the grey negation key, $\boxed{-}$, bottom row, column four and the subtraction key, $\boxed{-}$, row eight, right column.

# 2 Lists

1. **Lists L1, L2, L3, L4, L5, L6 are not displayed**. - To display all lists L1 through L6, press $\boxed{\text{STAT}}$, selecting 5: SetUpEditor and press $\boxed{\text{ENTER}}$ to display the command on the homescreen and press $\boxed{\text{ENTER}}$ to execute the command.

2. **Clearing a list.** - Use the up arrow key, $\boxed{\blacktriangle}$, to move the cursor to the column heading, not the first number of the list. Press $\boxed{\text{CLEAR}}\boxed{\text{ENTER}}$ to clear all numbers from the list, moving the cursor to the first row of the list.

3. **Left out a data value?** - Use the arrow keys, to move to the number just after where the missing data value should go. Press $\boxed{\text{2nd}}\boxed{\text{INS}}$, located on the second row, column three, and a space will open on the list. Enter the missing data value.

# 3 Graphing

1. **Screen Blank?** - Press $\boxed{\text{ZOOM}}$, 9: ZoomStat to adjust the window to show the stat plot.

2. **Extra lines?** - Press $\boxed{\text{Y=}}$. Clear any equations showing on Y1=, Y2=, etc. Press $\boxed{\text{ZOOM}}$, 9: ZoomStat to display the stat plot.

# 4 Correlation Coefficient

1. **Missing correlation coefficient?** - Press [2nd][CATALOG], using the down arrow key, [▼], locating DiagnosticOn. Press [ENTER] to select the command, placing the command on the homescreen. Press [ENTER] to execute the command. Press [2nd][ENTER], [2nd][ENTER] to recall your original instruction and press [ENTER] to execute the command once again.