TI-83/84 MANUAL

for Moore, McCabe, and Craig's

Introduction to the Practice of Statistics
Sixth Edition

Patricia Humphrey
Georgia Southern University

W.H. Freeman and Company
New York

Copyright © 2009 by W.H. Freeman and Company

No part of this book may be reproduced by any mechanical, photographic, or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted, or otherwise copied for public or private use, without written permission from the publisher.

Printed in the United States of America

ISBN: 1-4292-1475-9

Contents

Preface v

CHAPTER 0
Introduction to TI Calculators 1
0.1 Key Differences Between Models 2
0.2 Keyboard and Notation 2
0.3 Setting the Mode 4
0.4 Screen Contrast and Battery Check 5
0.5 The TI-89 Titanium 5
0.6 Home Screen Calculations 6
0.7 Sharing Data 7
0.8 Working with Lists 8
0.9 Using the Supplied Datasets 12
0.10 Memory Management 13
0.11 Common Errors 14

CHAPTER 1
Looking at Data—Distributions 15

1.1 Displaying Distributions with Graphs 16
1.2 Describing Distributions with Numbers 21
1.3 Density Curves and Normal Distributions 24
1.4 Common Errors 30

CHAPTER 2
Looking at Data—Exploring Relationships 32

2.1 Scatterplots 33
2.2 Correlation 36
2.3 Least-Squares Regression 38
2.4 Cautions About Correlation and Regression 40
2.5 Common Errors 42

CHAPTER 3
Producing Data 43

3.1 First Steps 44
3.2 Design of Experiments 46
3.3 Sampling Design 48
3.4 Toward Statistical Inference 50

CHAPTER 4
Probability: The Study of Randomness **52**

4.1 Randomness **53**
4.2 Probability Models **54**
4.3 Random Variables **56**
4.4 Means and Variances of Random Variables **58**
4.5 General Probability **59**

CHAPTER 5
Sampling Distributions **63**

5.1 Sampling Distributions for Counts and Proportions **64**
5.2 Poisson Random Variables **67**
5.3 The Sampling Distribution of a Sample Mean **68**
5.4 Common Errors **71**

CHAPTER 6
Introduction to Inference **72**

6.1 Confidence Intervals with σ Known **73**
6.2 Tests of Significance **76**
6.3 Use and Abuse of Tests **78**
6.4 Power and Inference as a Decision **79**

CHAPTER 7
Inference for Distributions **81**

7.1 Inference for the Mean of a Population **82**
7.2 Comparing Two Means **87**
7.3 Optional Topics in Comparing Distributions **91**

CHAPTER 8
Inference for Proportions **93**

8.1 Inference for a Single Proportion **94**
8.2 Comparing Two Proportions **97**

CHAPTER 9
Inference for Two-Way Tables **101**

9.1 Data Analysis for Two-Way Tables **102**
9.2 Inference for Two-Way Tables **103**
9.3 Formulas and Models for Two–Way Tables **105**

9.4 Goodness of Fit 108

CHAPTER 10
Inference for Regession 110

10.1 Simple Linear Regression 111
10.2 More Detail about Simple Linear Regression 119

CHAPTER 11
Multiple Regression 122

11.1 Inference for Multiple Regression 123
11.2 A Case Study 128

CHAPTER 12
One-Way Analysis of Variance 131

12.1 Inference for One-Way Analysis of Variance 132
12.2 Comparing the Means 135

CHAPTER 13
Two-Way Analysis of Variance 137

13.1 Plotting Means 138
13.2 Inference for Two-Way ANOVA 138

CHAPTER 14
Bootstrap Methods and Permutation Tests 144

14.1 The Bootstrap Idea 145
14.2 First Steps in Using the Bootstrap 146
14.3 How Accurate Is a Bootstrap Distribution? 151
14.4 Bootstrap Confidence Intervals 152
14.5 Significance Testing Using Permutation Tests 156

CHAPTER 15
Nonparametric Tests 161

15.1 The Wilcoxon Rank Sum Test 162
15.2 The Wilcoxon Signed Rank Test 165
15.3 The Kruskal-Wallis Test 169

CHAPTER 16
Logistic Regression **172**

16.1 The Logistic Regression Model **173**
16.2 Inference for Logistic Regression **175**

CHAPTER 17
Statistics for Quality: Control and Capability **179**

17.1 Statistical Process Control **180**
17.2 Using Control Charts **183**
17.3 Process Capability Indexes **186**
17.4 Control Charts for Sample Proportions **188**

CHAPTER 18
Time Series Forecasting **191**

18.1 Trends and Seasons **192**
18.2 Time Series Models **195**

Index of Programs **200**

Exercises **204**

Preface

The study of statistics has become commonplace in a variety of disciplines, and the practice of statistics is no longer limited to specially trained statisticians. The work of agriculturists, biologists, economists, psychologists, sociologists, and many others now quite often relies on the proper use of statistical methods. However, it is probably safe to say that most practitioners have neither the time nor the inclination to perform the long, tedious calculations that are often necessary in statistical inference. Fortunately there are now software packages and calculators that can perform many of these calculations in an instant, thus freeing the user to spend valuable time on methods and conclusions rather than on computation.

With their built-in statistical features, Texas Instruments' graphing calculators have revolutionized the teaching of statistics. Students and teachers have instant access to most commonly used statistical procedures. Advanced techniques can be programmed into the calculator which then make it (almost) as powerful as, but much more convenient than, common statistical software packages.

This manual serves as a companion to your W. H. Freeman Introductory Statistics text. Examples either taken from the text, or similar to those in the text, are worked using either the built-in TI calculator functions or programs specially written for the calculator. The tremendous capabilities and usefulness of TI calculators are demonstrated throughout. It is hoped that students, teachers, and practitioners of statistics will continue to make use of these capabilities, and that readers will find this manual helpful.

Programs

All codes and instructions for the programs are provided in the manual; however, they can be downloaded directly from the author's website at **http://math.georgiasouthern.edu/~phumphre/TIpgms/** or from your text's portal.

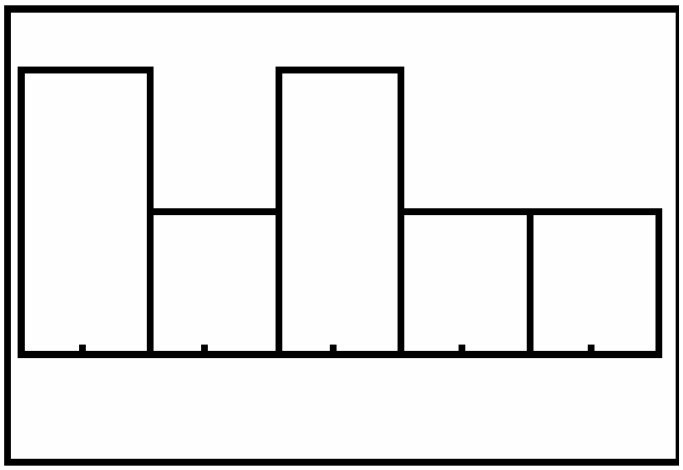
Acknowledgments

My thanks go to W. H. Freeman and Company for giving me the opportunity to revise the manual to accompany their various texts. Special thanks go to Ruth Baruth and editorial assistant Jennifer Albanese for her organization and help in keeping me on schedule. As always, my sincere gratitude goes to Professor Moore and his coauthors for providing educators and students with an excellent text for studying the practice of statistics.

Patricia B. Humphrey
Department of Mathematical Sciences
Georgia Southern University
Statesboro, GA 30460-8093

CHAPTER

0



Introduction to TI Calculators

0.1	Key Differences Between Models
0.2	Keyboard and Notation
0.3	Setting the Mode
0.4	Screen Contrast and Battery Check
0.5	The TI-89 Titanium
0.6	Home Screen Calculations
0.7	Sharing Data
0.8	Working with Lists
0.9	Using the Supplied Datasets
0.10	Memory Management
0.11	Common Errors

Introduction

In this chapter we introduce our calculator companion by giving an overview of Texas Instruments' graphing calculators: the TI-83+, -84+, and -89. Read this chapter carefully in order to familiarize yourself with the keys and menus most utilized in this manual. These skills include Home screen calculations and saving and editing lists of data in the STAT(istics) editor.

0.1 Key Differences Between Models

All calculators in the TI-83/84/89 series have built-in statistical capabilities. Although a few statistical functions are "native" on the TI-89, most of the topics covered in a normal Statistics course require downloading the Texas Instruments Statistics with List Editor application which is free. Download requires the TI-Connect cable. See the web page http://education.ti.com/us/product/tech/89/apps/appslist.html for more information. This manual assumes the statistics application has been loaded on the calculator. If you have the newer TI-89 Titanium edition, the statistics application comes pre-loaded, and the TI-Connect cable is included with the calculator.

The TI-83 and -84 series calculators are essentially keystroke-for-keystroke compatible; however, the 84 does have some additional capabilities (some additional statistical distributions and tests, for example) with the latest version of the operating system, version 2.41 which is also available for download at education.ti.com.

There are some major differences in calculator operation and menu systems that will in some cases necessitate separate discussions of procedures for the TI-83/84 and TI-89 calculators. Some of these will become apparent in the next section. Not only are there differences between the three series, but there is also a difference in operation between the TI-89 and the TI-89 Titanium edition. When a regular TI-89 is turned on, the user is on the "home screen" similar to that for the TI-83 and -84. On the Titanium, all "functions" on the calculator are essentially applications; when the Titanium edition is first turned on, one must scroll using the arrow keys to locate the desired application, but we'll say more about this later.

0.2 Keyboard and Notation

All TI keyboards have 5 columns and 10 rows of keys. This may seem like a lot, but the best way to familiarize yourself with the keyboard is to actually work with the calculator and learn out of necessity. The keyboard layout is identical on the 83+ and 84+. The layout of the 89 (and 89 Titanium) keyboard is similar, but some functions have been relocated. You will find the following keys among the most useful and thus they are found in prominent positions on the keyboard.

- The cursor control keys [◄], [►], [▲] and [▼] are located toward the upper right of your keyboard. These keys allow you to move the cursor on your screen in the direction the arrow indicates.
- The [Y=] key is the leftmost key on the top row of 83 and 84 keyboards. It is utilized more in other types of mathematics courses (such as algebra) than in a statistics course; however you will use the [2nd] function above it quite often. This is the STAT PLOT menu on the -83/84 series. We will discuss [2nd] functions shortly. On TI-89 calculators, the Y= application is accessed by pressing [◆][F1]. The STAT PLOT menu on the TI-89 series is found inside the Statistics application.
- The [ON] key is in the bottom left of the keyboard. Its function is self-explanatory. To turn the calculator off, press [2nd][ON].

- The [ENTER] key is in the bottom right of the keyboard. You will usually need to press this key in order to have the calculator do what you have instructed it to do with your preceding keystrokes.
- The [GRAPH] key is in the upper right of the keyboard. On the TI-89, GRAPH is [◆][F3].

As mentioned briefly above, most keys on the keyboard have more than one function. The primary function is marked on the key itself and the alternative functions are marked in color above the key. Actual color depends on the calculator model. Next we describe how to engage the functions that appear in color.

The [2nd] key

The color of this key varies with calculator model, but in all models it is the leftmost key on the second row. If you wish to engage a function that appears in the corresponding color above a key, you must first press the [2nd] key. You will know the second key is engaged when the cursor on your screen changes to a blinking ↑. As an example, on a TI-83 or -84 if you wish to call the STAT PLOTS menu, which is in color above the [Y=] key, you will press [2nd] [Y=].

The [ALPHA] key

You will also see characters appearing in a second color above keys that are mostly letters of the alphabet. There are some situations in which you will wish to name variables or lists and in doing so you will need to type the names. If you wish to type a letter on the screen you must first press the [ALPHA] key. The color of this key depends on the model of calculator and corresponds to the color of the letters above the keys. You will know the [ALPHA] key has been engaged when the cursor on the screen turns into a blinking A. After pressing the [ALPHA] key you should press the under where the letter appears. As an example if you wish to type the letter E on an 83 or 84, press [ALPHA] [SIN] (because E is above [SIN]. To get the same letter E on an 89, press [ALPHA][÷].

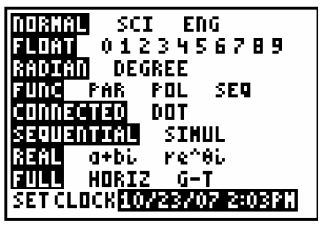
Note: If you have a sequence of letters to type, you will want to press [2nd] [ALPHA]. This will engage the colored function above the [ALPHA] key which is the A-LOCK function. It locks the calculator into the Alpha mode, so that you can repeatedly press keys and get the alpha character for each. Otherwise, you would have to press [ALPHA] before each letter. Press [ALPHA] again to release the calculator from the A-LOCK mode.

Some general keyboard patterns and important keys

1. The top row on 83's and 84's is for plotting and graphing. On 89's these functions are also on the top row, but are accessed by preceding the desired function with [◆].
2. The second row from the top has the important QUIT function ([2nd] [MODE] on 83's and 84's, [2nd][ESC] on 89's). On 83's and 84's it also contains the keys useful for editing ([DEL], [2nd] [DEL] (INS), [◄], [►], [▲] and [▼]). INS and DEL on 89's are both combination commands: INS is [2nd][←] and DEL is [◆][←].
3. The [MATH] key in the first column on 83's and 84's leads to a set of menus of mathematical functions. Several other mathematical functions (like [x^2]) have keys in the first column. On a TI-89, [2nd][5] leads to the Math menu.

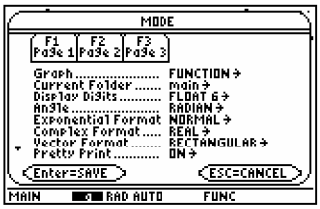
4. The keys for arithmetic operations are in the rightmost column (÷ × − +).
   Note: When displaying input, the ÷ shows as /, and the × shows as *. On both 89 models, when the command is transferred to the display area the * is replaced with a · and division looks like a fraction.
5. The STAT key will be basic to this course. Submenus from this allow editing of lists, computation of statistics, and calculations for confidence intervals and statistical tests. On 89's with the statistics application, one starts the application using the key sequences ♦APPS and selecting the Statistics application. On the 89 Titanium, quit the current application (2nd ESC) and locate the Stats/ListEd application, and press ENTER to start the application. On 83's and 84's the second function of the STAT key is LIST. This key and its submenus allow one to access named lists and perform list operations and mathematics.
6. The VARS key on 83's and 84's allows one to access named variables. On TI-89's this key is 2nd − which is named [VAR-LINK], and is used to access names of lists and variables.
7. 2nd VARS calls the distributions (Distr) menu. This is used for many probability calculations. To get this menu on a TI-89, press F5 from within the Stats/ListEd application.
8. The . key is located directly above the 7 key on 83's and 84's; on 89's it is above the 9 key. It is used quite often for grouping and separating parameters of commands.
9. The STO▶ key is used for storing values. It is located near the bottom left of the keyboard directly above the ON key on all the calculators. It appears as a → on the display screen.
10. The (-) key on the bottom row (to the left of ENTER) is the key used to denote *negative* numbers. It differs from the subtraction key −.
    Note: The (-) shows as ⁻ on the screen, smaller and higher than the subtraction sign.
11. On TI-89 models, switching back and forth between two apps is easiest done by pressing 2nd APPS.
12. The HOME key on TI-89 models switches the application from wherever you are to the home screen immediately.

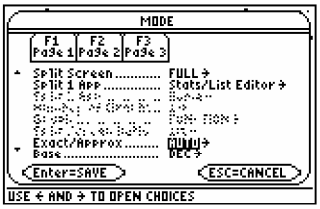
0.3 Setting the Mode

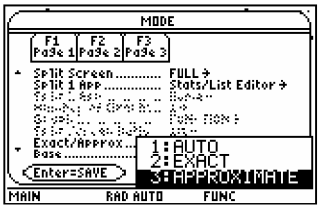


If your answers do not show as many decimal places as the ones shown, or if you have difficulty matching any other output, check your MODE settings. On an 83 or 84, press the MODE key (second row, second column). If your calculator has been used previously by you or someone else the highlighted choices may differ. If your screen has different highlighted choices use the ▲ and ▼ keys to go to each row with a different choice and press ENTER when the blinking cursor is on the first choice in each row. This will move the highlight to the first choice in each row. Continue until your screen looks like the one at right. Press 2nd MODE (QUIT) to return to the Home Screen.

On TI-89's, the default is to give "exact" answers. For statistical calculations, we will want decimal approximations. To set this, press MODE. Press F2 to proceed to the second page of settings, then arrow to Exact/Approx and use the right and down arrows to change the setting to 3:Approximate. Press ENTER to complete the set-up. The sequence is shown below.







0.4 Screen Contrast and Battery Check

To increase the contrast, press and release the 2nd key and hold down the ▲ key. You will see the contrast increasing. There will be a number in the upper-left corner of the screen that increases from 0 (lightest) to 9 (darkest).

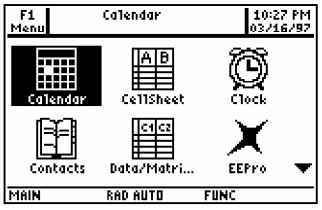
To decrease the contrast, press and release the 2nd key and hold down the ▼ key. You will see the contrast decreasing. The number in the upper-left corner of the screen will decrease as you hold. The lightest setting may appear as a blank screen. If this occurs, simply follow the instructions for increasing the contrast, and your display will reappear.

When the batteries are low, the display begins to dim (especially during calculations) and you must adjust to a higher contrast setting than you normally use. If you have to set the contrast setting to 9, you will soon need to replace the four AAA batteries. With newer versions of the operating system, your calculator will display a low-battery message to warn you when it is time to change the batteries. After you change batteries, you will need to readjust your contrast as explained above.

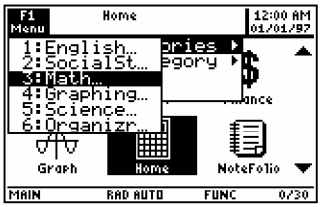
Note: It is important to turn off your calculator and change the batteries as soon as you see the "low battery" message in order to avoid loss of your data or corruption of calculator memory. Change batteries as quickly as possible. Failure to do so may result in the calculator resetting memory to factory defaults (losing any data or options which have been set).

0.5 The TI-89 Titanium

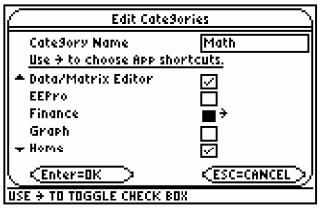
On the TI-89 Titanium, most important functions that on other calculators are accessed by keystrokes, are applications (Apps). When the calculator is first turned on, you will be presented with a graphical menu of these applications, as at right. Paging through the screen to find the one you want can be tiresome and time consuming. There is a way to customize this screen so that you only see those applications you want to see.



Press F1. Press the right arrow key to expand menu selection 1:Edit Categories. You will be presented with a list of possible categories. Press 3 to select option 3:Math.



On this screen, use the down arrow to page through the list of applications. When you find one you want to be displayed, press the right arrow key to place a checkmark in the box. The screen at right shows that the Data/Matrix Editor and the Home screen have been selected. For this statistics course, you will want these applications, along with the Stats/List Editor and Y= applications. Press ENTER when you have finished making your selections.



On this calculator, pressing 2nd ESC (Quit) will return you to the applications selection screen.

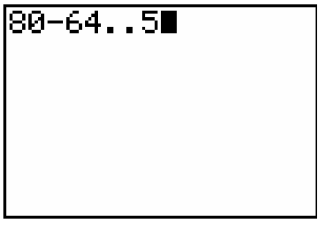
0.6 Home Screen Calculations

The following example illustrates some techniques which will be useful in performing home screen computations. We also point out the importance of correctly using parentheses in calculations.

Example 0.1: $\dfrac{70-64.5}{2.7}$

We will calculate the value in two ways. In doing so, we will intentionally make a mistake to show you how to correct errors using the DEL key. We also discuss the Ans and Last Entry features.

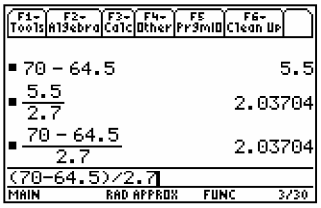
Type 80-64..5 (two intentional mistakes). To correct these, use the ◄ cursor key to move backward until your cursor is blinking on one of the double decimal points. Press DEL (on an 89, either ♦← or position the cursor to the right of the character to be deleted and press ←) and the duplicate decimal point will be deleted. Now press ◄ until the cursor is blinking on the 8. Type a 7, and it will replace the incorrect 8. On an 89, move the cursor to the right of the error, press ← and then type the correct 8. Press ENTER for the numerator difference of 5.5 as shown in the top of the screen below.



Press ÷. (Note that "Ans/" appears on the screen). Type 2.7 and press ENTER for the result of 2.037.

Note: Ans represents the last result of a calculation that was displayed alone and right-justified on the Home screen. Pressing ÷ without first typing a value called for something to be divided, so Ans was supplied.

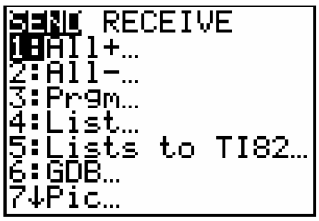
To do the calculation in one step, press 2nd ENTER. This calls the "last entry" to the screen (in this case Ans/2.7). Press 2nd ENTER again to get back to 70-64.5. Press the ▲ key to move to the front of the line. On an 89, press 2nd◄.



On an 83 or 84, you will need to press [2nd] [DEL] (for INS); 89's are always in insert mode. You will see a blinking underline cursor. Type [(] to insert a left parenthesis before the 7. Press [▼] ([2nd][▶] on an 89) to jump to the end of the line. Type [)] [÷] 2.7 to see the result. Press [ENTER] for the same result as before.

0.7 Sharing Data

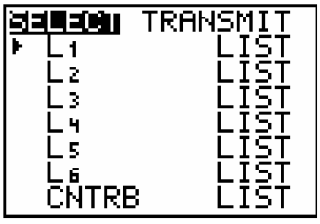
Sharing data between calculators (TI-83/84)



Data and programs may be shared between calculators using the communications cable which is supplied. The TI-83 and 84 series can share any TI-83/84 information with the exception of flash applications and their associated variables.

On the TI-83+, the I/O port is at the base of the calculator. On the TI-84+, you can use either the USB port or the I/O port on the top to link to another 84 series calculator. To link to an 83 series, you must use the I/O port. Connect the appropriate cable to the ports. On both calculators, press [2nd] [X,T,Θ,n] to activate the LINK menu.

On the receiving calculator, press [▶] to highlight RECEIVE, then press [ENTER]. The calculator will display the message "Waiting..." The rolling cursor on the upper right indicates the calculator is working.

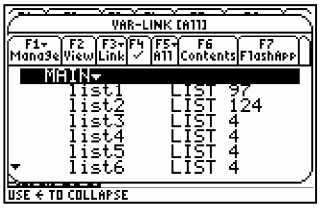


On the sending calculator, use the arrow keys to select the type of information to send. For example sending lists, either arrow to 4:List and press [ENTER] or press [4]. The screen at right will be shown.

To select items to send, move the cursor to the item, press [ENTER] to select it. After selecting all items to send, press [▶] to highlight TRANSMIT, then press [ENTER].

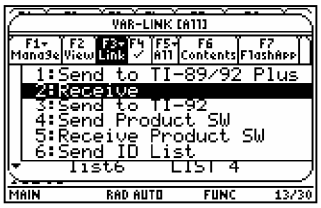
Sorry, TI-83's and 84's cannot communicate with TI-89's.

Sharing data between calculators (TI-89 series)



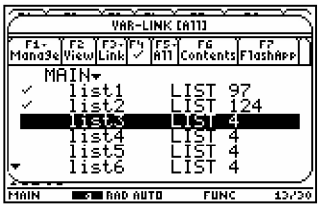
Data and programs may be shared between calculators using the communications cable which is supplied. The TI-89 can only communicate with the other TI-89's and TI-92s.

Connect the supplied cable to the port at the base of each calculator. On both calculators press [2nd] [−] to activate the VAR-LINK menu.

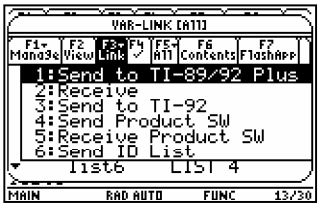


On the receiving calculator, press [F3] to select Link, [▼] to highlight Receive, then press [ENTER]. The screen reverts to main VAR-LINK Menu with a "Waiting to receive" message at the bottom.

On the sending calculator, use the arrow keys to select the item to send, then press [F4] to "check" the item. The example at right will send `list1` and `list2`.



Now press [F3] to select Link, then [ENTER] to select menu choice `1:Send to TI-89/92Plus` which is highlighted by default.



An analogous procedure can be used to send applications between calculators. Applications (such as the Statistics with List Editor) are selected from the [F7] `FlashApp` menu (press [2nd][F2] for [F7].)

Sharing data between the calculator and a computer

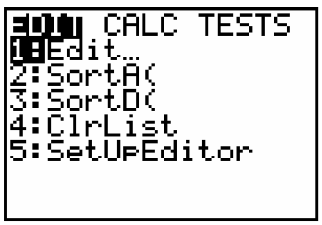
Data lists, screen shots, and programs may be shared between the calculator and either Microsoft Windows or Macintosh computers using a special cable and TI-Connect or software. This can be helpful since all data sets in the text are included on the CD and companion website for your text. Cables for the TI-83+ are available for either serial or USB ports; they can be found through many outlets such as OfficeMax, or Amazon.com. The software can be downloaded free through the Texas Instruments website at education.ti.com. The needed USB cable and computer software are included with the TI-84 and -89 models. See the section on using supplied data sets for more information.

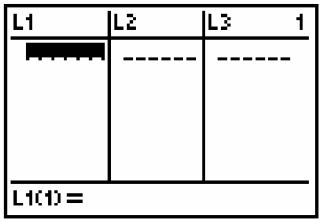
0.8 Working with Lists

The basic building blocks of any statistical analysis are lists of data. Before doing any statistics plot or analysis the data must be entered into the calculator. The calculator has six lists available in the statistics editor; these are `L1` through `L6` (`list1` through `list6` on an 89). Other lists can be added if desired. The number of lists is only limited by the memory size.

The statistics editor

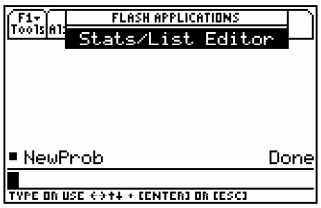
On a TI-83/84, press [STAT], `1:Edit...` will be highlighted. Press [ENTER] to select this function.

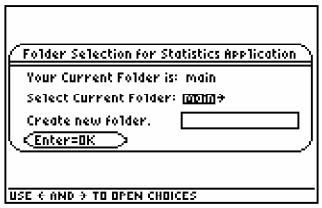


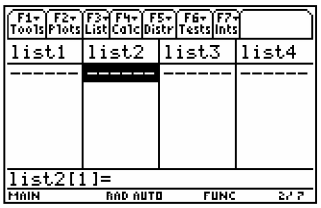


To enter data, simply use the right or left arrows to select a list, then type the entries in the list, following each value with [ENTER]. Note that it's not necessary to type any trailing zeros. They won't even be seen unless decimal places (found with the [MODE] key) have been set to some fixed number.

On a TI-89, press [◆][APPS] followed by the selection of the Statistics/List Editor Flash Application followed by [ENTER]. On the Titanium edition, select the Statistics/List Editor from the main application menu screen. If this is the first time the editor has been accessed since the calculator has been turned on, you will be prompted for a data folder as in the middle screen. The default folder is `main`. Press [ENTER] to select `main` as the current folder, or press ⊙ to allow a new folder to be created. To enter a new folder name, arrow to the entry block and type the name of the new folder. Your lists will be stored in the new folder and it will be set to default. To change folders, press ⊙ to select a folder. Then press [ENTER] to proceed to the list editor. If the editor has been used since the calculator has been turned on, pressing [ENTER] to select the application will automatically open the editor.



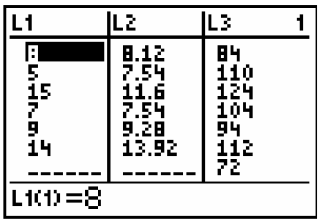




One word of advice: Most lists of data in texts are entered across the page in order to save space. Don't think that just because there are four (or more!) columns of data they belong in four (or more) lists. Data that belongs to a single variable always belongs in a single list.

Entering data into the STAT editor

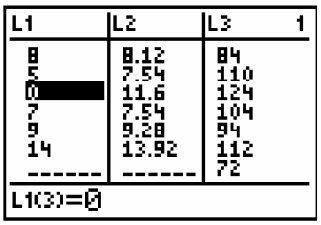
With the cursor at the first row of `L1`, type 8 and press [ENTER]. The cursor moves down one row. Type 5 followed by [ENTER] and the 5 will be pasted into the second row of the list. Continue with 15, 7, 9, and 14 as seen at right.



Correcting mistakes with DEL and INS

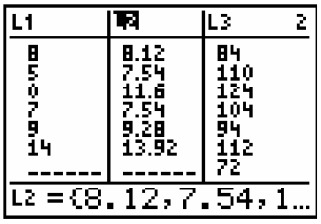
In the screen above, we can delete the 15 by using the ▲ key until it is highlighted and then pressing [DEL] ([◆][←] on an 89).

To insert a 12 above the 7 move the cursor to the 7 then press [2nd] [DEL] ([2nd][←] on an 89) (to choose `INS` or insert mode). Note a 0 was inserted where you wanted the 12 to go. Just type over the place-holding 0 with the value you want.



Clearing lists without leaving the STAT editor

Suppose you wish to clear a list, say L2, while you are still in the STAT Editor. You should use the cursor to highlight the name of the list at the top. With the name highlighted, press [CLEAR] and you will see this. Press [ENTER] and the contents of the list will be cleared. *Make sure not to press* [DEL] or the list will be deleted entirely and you will have to use SetUpEditor as described below to retrieve it.

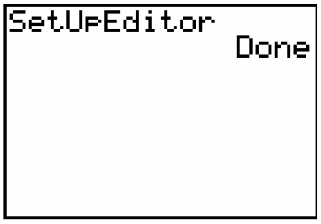


Deleting a list from the STAT editor

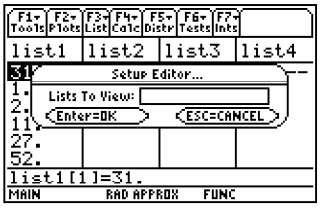
If you wish to delete a list from your STAT Editor, simply highlight the list name and press [DEL]. The name and the data are gone from the Editor but not from the memory. To recover a list inadvertently deleted, use SetUpEditor as described below.

SetUpEditor

Setting up the editor will remove unwanted lists from view. It also will recover lists that have inadvertently been deleted. On an 83 or 84, if you want the STAT Editor to be restored to its original condition (with lists L1 to L6 only), press [STAT] [5] [ENTER]. Often students find this necessary because they have inadvertently deleted one of the original lists.

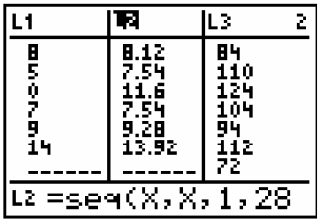


On an 89, in the Statistics Editor, press [F1] (Tools), then select option 3:Setup Editor. You will see the screen at right. Leave the box empty and press [ENTER] to return to the six default lists.

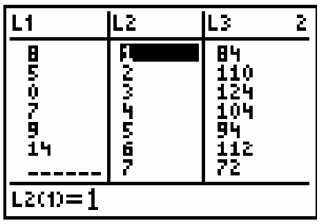


Generating a sequence of numbers in a list

From time to time one may want to enter a list of sequenced values (years for example in making a time-series plot). It is certainly possible (but tedious) to enter the entire sequence just as one would enter normal data. There is an easier option, however. Use cursor control keys to highlight L1 in the top line. Press [2nd] [STAT] [▸] [5]. You are choosing the LIST menu and then choosing the OPS submenu. From the OPS submenu you choose option 5 which is seq(.



This has been pasted onto the bottom line of the screen. Type in the rest so that you have seq(X,X,1,28. Press [ENTER] and the sequence of integers from 1 to 28 will be pasted into L2 as in my screen. Find the X on the [X,T,Θ,n] key on an 83 or 84, on 89's it has its own key. Notice that it was not necessary to clear the list first.

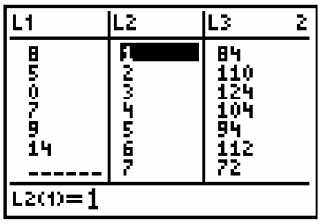


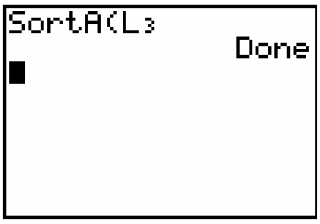
On an 89, the procedure is analogous, but access the LIST OPS menu by pressing F3 2, then select option 5.

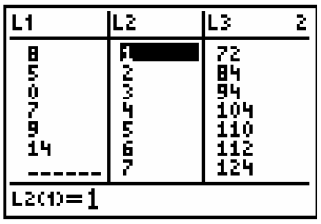
Note: To quickly check the values on a multi-screen list you can press the green ALPHA key followed by either the ▲ or ▼ key. This will allow you to jump up or down from one page (screen) to another. The green arrows on the keyboard near the ▲ and ▼ keys are there to remind you of this capability. On a TI-89, instead of the ALPHA key, press ◆.

Sorting lists (TI-83/84)

Lists may be sorted in either ascending (smallest to largest value) or descending order. The resulting list will replace the original list. On the main STAT menu select either 2:SortA(or 3:SortD(. The command will be transferred to the home screen. Enter the name of the list to be sorted (2nd n where n is the number of the list). Execute the command by pressing ENTER. The example below sorts list L3.

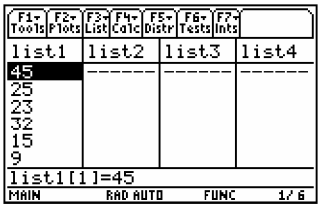


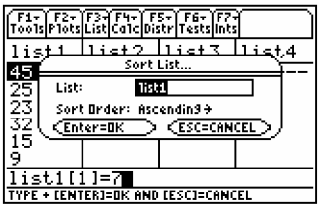


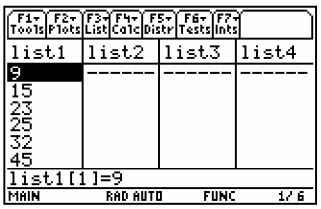


Sorting lists (TI-89)

While in the List Editor, press F3 for the List menu, press ▼ followed by ENTER or 2 to select List Ops, then press ENTER to Select 1:Sort List, 2nd − (Var-Link) and use the arrows to select the list to be sorted. Press the right arrow if necessary to change the sort order (use ▶ to activate the menu choices). Press ENTER to carry out the command.

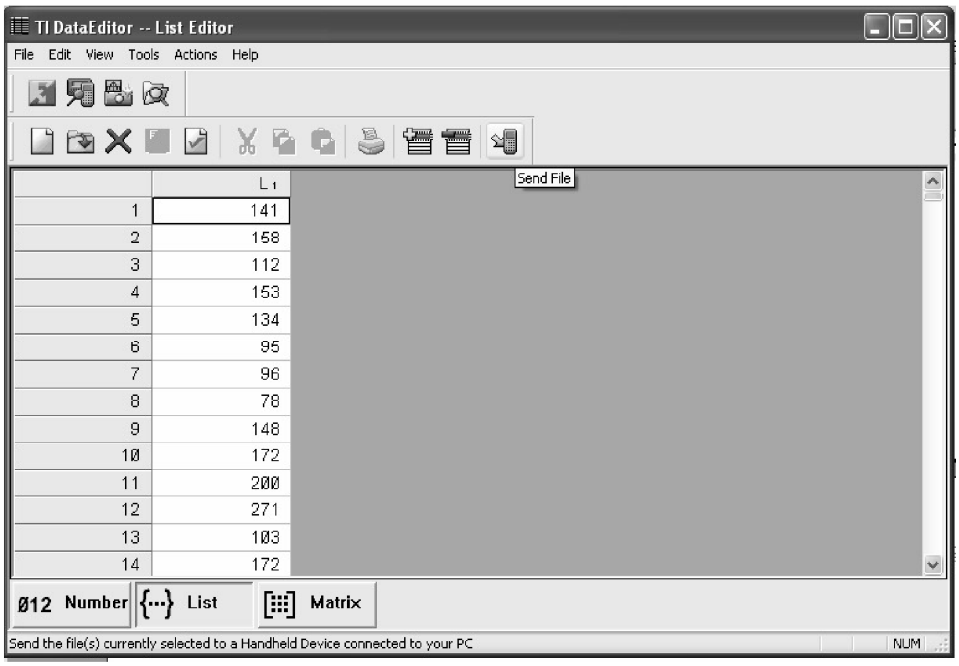






0.9 Using the Supplied Datasets

If you have the additional Connect cable and TI-Connect software, you can load the datasets for all examples, exercises, and tables in the text easily without needing to retype all the numbers. Datasets are also on the text's accompanying web site. If you're using the CD, insert it in the drive — it should automatically open (you may get a warning about "dangerous content" — and simply click to continue. Click on the datasets link, then select PC or Mac TI-83 format. You may at this point want to copy the folder to your desktop for easier future use. Click on the folder to open it. Select the file you wish to use — names beginning with eg are examples, ex are exercises, and ta are tables. In the example below, I have opened a file from Chapter 1 of the *Basic Practice of Statistics* for exercise 11 about blood glucose levels. Notice that its heading is L1.



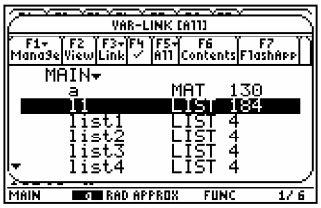
To transfer the file to the calculator, press the Send File icon (at top right). The transfer should start automatically, after the software locates your calculator. You may see a warning like the one below:

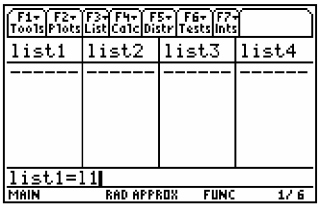


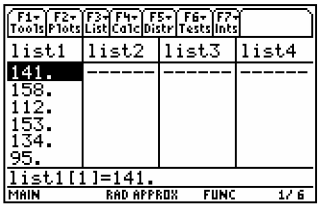
Press [ENTER] to replace the current contents of list L1. If you want to save the current contents of L1, click to rename and send the data to another list, or select Cancel to abort the transfer.

Additional TI-89 step

With these calculators, the lists will default to being named L1, L2, etc in the main folder. They can be used (and accessed) with [VAR-LINK] just like the default lists for graphs and other calculations. If you want to be able to look at them in the statistics editor, highlight the desired statistics editor list name, then use [2nd][−] ([VAR-LINK]) and move the cursor to the desired list name. Press [ENTER] to select the name, then [ENTER] again to fill the list.







0.10 Memory Management

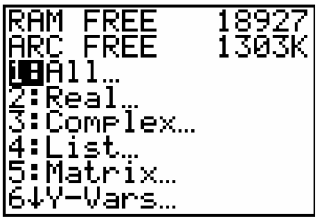
Too many applications loaded or lists in active memory can overload the calculator. Just as a computer disk can be filled up, so can memory on the calculator. The TI calculators have two types of memory – RAM and archival. If you use the applications supplied on the CD (on a TI-83/84), these are loaded into archival memory. Active lists are in RAM.

TI-83/84 procedure



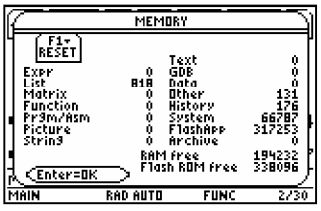
To find out the current free memory status, press [2nd][+] (MEM) and select option 2:Mem Mgmt/Del.

The screen at right shows my calculator currently has 18,927 bytes (characters) of free RAM and 1303Kilobytes of free archival memory.



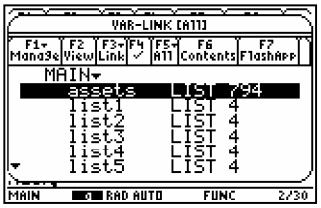
If you need to free some memory, decide the type. If you want to delete some lists, for example, select 4:List. Move the cursor to the lists you wish to delete and press [DEL] for each one. This can also be done for any applications you no longer need from previous chapters, but use choice 5:Archive to access the list of archived applications.

TI-89 series procedure

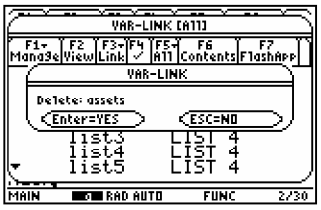


To find out the current free memory status, press [2nd][6]. The screen at right shows I currently have 194,232 free bytes of RAM and 338096 free bytes of Flash memory free. Pressing [F1] here will reset

RAM, Flash, or all memory to either totally blank or factory default settings. I do not recommend either of these options under normal circumstances. Resetting Flash, for example, would erase the Statistics Flash application, which would then need to be reloaded.



To delete lists that are no longer needed, press 2nd - (VAR-LINK). Move the cursor to highlight the list to be deleted, then press F1. Press ENTER to select option 1:Delete.

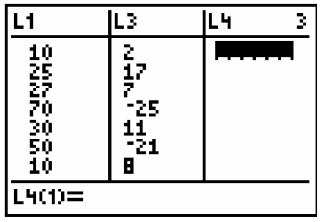


You will be prompted to verify that the selected item is to be deleted. Press ENTER to confirm the deletion, or ESC to cancel. You can continue this process to delete all unneeded lists.

0.11 Common Errors

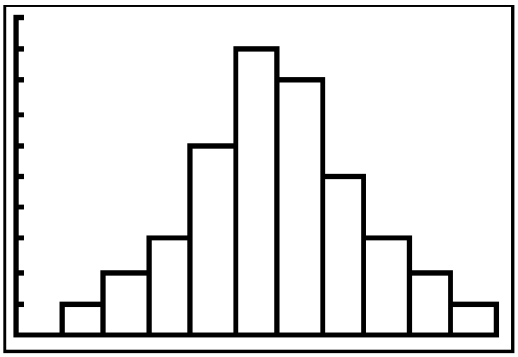
Why is my list missing?

By far the most common error, aside from typographical errors is improper deletion of lists. When lists seem to be "missing" the user has pressed DEL rather than CLEAR in attempting to erase a list. Believe it or not, the data and the list are still in memory. To reclaim the missing list press STAT and select choice 5:SetUpEditor followed by ENTER to execute the command. Upon return to the Editor, the missing list will be displayed.



CHAPTER

1



Looking at Data—Distributions

1.1	Displaying Distributions with Graphs
1.2	Describing Distributions with Numbers
1.3	Density Curves and Normal Distributions
1.4	Common Errors

Introduction

In this chapter, we use the TI calculators to view data sets. We first show how to make bar graphs, histograms, and time plots. Then we use the calculator to compute basic statistics, such as the mean, median, and standard deviation, and show how to view data further with boxplots. Lastly, we use the TI-83 Plus for calculations involving normal distributions.

1.1 Displaying Distributions with Graphs

We start by using the calculator to graph data sets. In this section, we will use the STAT EDIT screen to enter data into lists and use the STAT PLOT menu to create bar charts, histograms, and time plots.

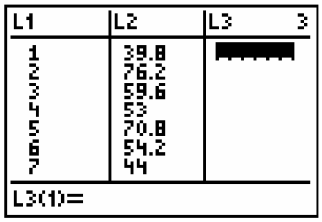
Throughout the manual, we will be working with data that is entered into lists L1 through L6 on the TI-83/84 (list1 through list6 on an 89 model). These lists can be found in the STAT EDIT screen. A list should be cleared before entering new data into it.

Example 1.1 Women's Degrees: A Bar Graph of Categorical Data. TI calculators cannot make true bar graphs, since all quantities entered on the STAT EDIT screen must be numeric. Also, proper bar graphs should have the bars separated. These calculators can however, guide you in creating a bar graph. Here are the percents of women among students seeking various graduate and professional degrees during the 1999–2000 academic year.

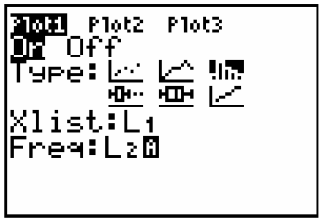
Degree	Percent female
MBA	39.8
MAE	76.2
Other MA	59.6
Other MS	53.0
Ed.D.	70.8
Other Ph.D.	54.2
MD	44.0
Law	50.2
Theology	20.2

We would like first to make a bar graph of the data.

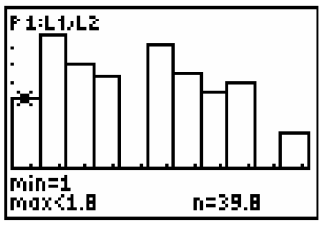
TI-83/84 Solution. First, label the nine categories as 1–9 and enter these values into list L1, then enter the percents into list L2.



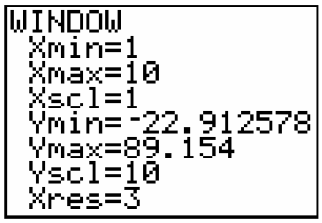
Press 2nd Y= (STAT PLOT). Press ENTER to select Plot1. Our plot will use the histogram plot type, and we will put the category numbers that were entered in L1 on the *X* axis, and the percentages on the *Y* axis, so our plot definition screen is at right. At this point, it is good practice to check that no functions are defined on the Y= screen itself (if so, press CLEAR to erase them) and that the other two plots are "turned off." *TI calculators try to plot everything they know of at once. This can result in error messages and other junk!*



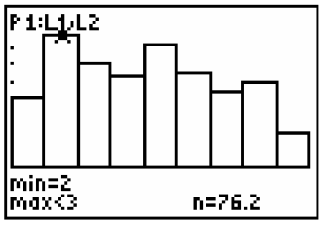
For most plots, pressing [ZOOM][9] (ZoomStat) will suffice. However, for histograms and bar charts this usually does not give a reasonable picture. At right is my ZoomStat graph. I have pressed [TRACE] to display the first bar. Notice that the bar's x-values go from 1 to 1.8. We need intervals of width 1 for our bar chart.



Press [WINDOW]. This allows us to control the values that display on a graph. If you look at the graph above, notice that the endpoint of the first interval (1.8) is not included. We need intervals of length 1. All we need to do here is change Xmax (the largest X value that displays to 10 (for consistency) and the bar width Xscl to 1. Don't worry about changing Ymin and Ymax. Ymin is negative so that the [TRACE] information does not obscure the graph.

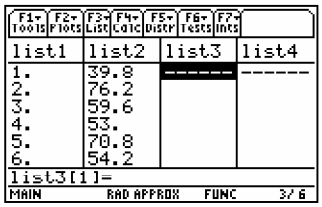


Press [GRAPH] to display the rescaled graph. *When you have changed a window, always press* [GRAPH]. *Pressing* [ZOOM][9] *will revert to the default scaling.* Here is the completed graph. If you are copying this onto paper, don't forget to give proper labels to the categories and separate the bars.

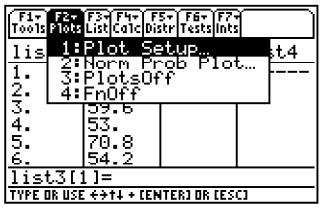


TI-89 Solution

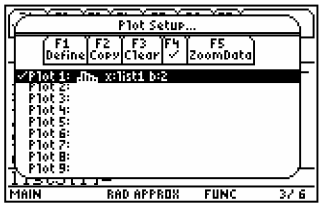
Press [◆][APPS], move the cursor to highlight the Statistics/List Editor Application, and press [ENTER] to get to the list editor. Here I have entered the data into list1 (the category numbers) and list2 (the percents).



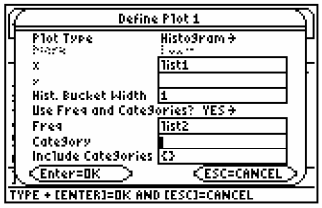
Plot Definition on TI-89 calculators is somewhat different from the 83/84models. Press [F2] (Plots) then [ENTER] to select option 1:Plot Setup.



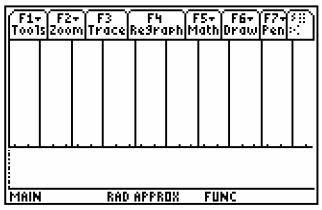
TI-83/84 calculators can have up to 3 plots defined at once, the 89 can have as many as 9. Press [ENTER] to select Plot1 to be defined. *TI calculators try to graph everything they possibly can at once.* This can lead to error messages and other types of "junk" on your graph. If any other plots have been defined (and you wish to preserve them) move the highlight to the plot line and press [F4] to uncheck them so they won't try to be displayed. Similarly, you should check that the [Y=] function editor (press [◆][F1]) is cleared.



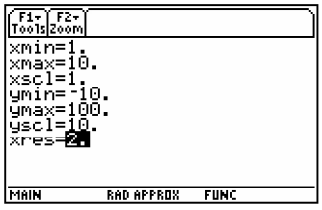
Press F1 to select Plot 1 to be defined. The plot type is a histogram, our category "labels" were stored in list1. TI-89 calculators must have the bar (bucket) with explicitly defined. Here, I have set it to one. Since we have another list of the frequencies of each category, I have set Use Freq and Categories to YES and specified that the frequencies are in list2. Remember, press F2 - ([VAR-LINK]) to locate the list names. Press ENTER to finish the plot definition. You will be returned to the Plot Setup screen.



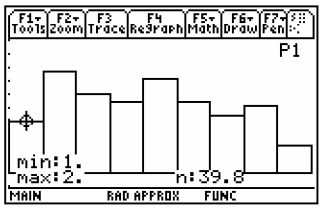
If I press F5 (ZoomData) at this point, this is the graph I see. TI-89 calculators (especially on bar charts like these and histograms also) need some extra help in setting their windows because of the tabs at the top of the screen. Press ◆F2 (WINDOW).



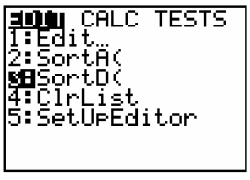
At right, I've changed the window settings. The minimum (xmin) and maximum (xmax) *x*-axis values to display have been set to 1 and 10 because our lowest category number was 1 and the bars will have width 1. ymin and ymax have been changed as well to allow the full bar graph to display. We need a "roomy" ymax so the tabs don't hide the top of the plot.

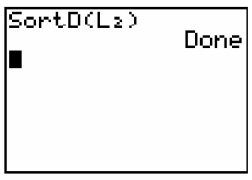


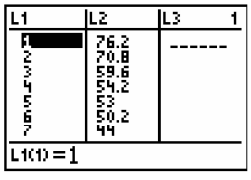
Press ◆F3 to display the graph. Press F3 (Trace) to display the values shown by each bar. The upper end of each bar (2 in the first bar highlighted) is not included in the bar. This will become important in histograms. In copying your graph onto paper, don't forget to use the proper category labels, and to separate the bars.

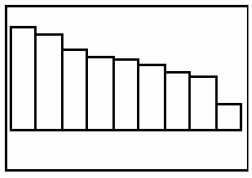


Example 1.2 Women's Degrees: A Pareto Chart of Categorical Data. Next, we make a *Pareto* chart, a bar chart with the bars ordered by height. To do so, we simply use the SortD(command from the STAT EDIT screen to sort the data in list L2 into descending order, then regraph using the same window settings as before by pressing GRAPH. On a TI-89, find the sort command on the F3 (List) option 2:Ops menu.









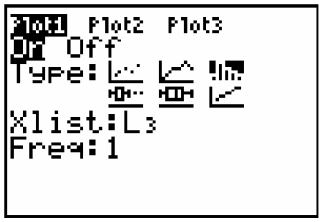
If copying this onto paper, be careful to match the bars with their correct labels (degrees)!

Example 1.3 The Density of the Earth: Making a Histogram. Make a histogram of Cavendish's measurements of the density of the Earth.

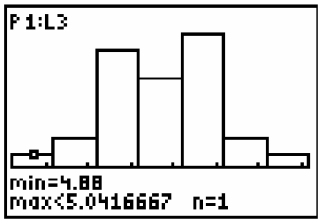
5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Solution. First, we enter the data into a list. Here we have entered the data in list L3.

First, define the plot as at right. In the bar chart above, we had a separate list of the frequencies of each type of degree, now each value represents one observation. To change the Freq: to 1, press [ALPHA] to change the cursor from A mode, or your typed [1] will be a Y.

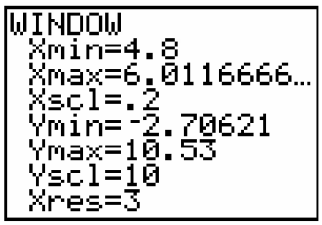


Press [ZOOM] [9] to display the default graph. Then press [TRACE] to display the actual bars and counts. As we saw with bar charts, the calculator needs some help setting a reasonable window. The length of this first bar (and succeeding bars) does not make logical sense. Your instructor (or text) may indicate what intervals should be used on any given problem, but in general we want these to make some type of logical sense so as to clearly convey the information, and use a reasonable number of bars.

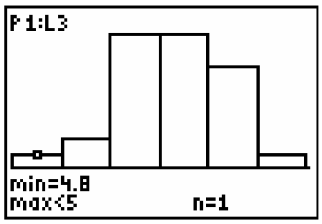


How many bars we need is matter of judgment, but usually we want between 5 and 20 bars. The old "rule of thumb" was to start by thinking of n (the sample size) divided by 5. Here, our n is 29, so about 6 bars would be reasonable. Looking at the minimum value, we need to make our minimum somewhat smaller, and find a bar width that is intuitive.

Here, I have chosen to set Xmin to 4.8 and use a bar width (Xscl) of .2. You normally will not need to change the Y settings unless a scale change makes the top of a bar disappear; in that case, increase Ymax.



Here is my final calculator graph. To copy it onto paper, use [TRACE] and the right arrow key to move across the bars. Each interval on the x-axis goes from the min to the max shown and contains the number of observations indicated by n= at the lower right. *Be careful the x-axis is a number line; do not display the values as "category labels!"*



This distribution is unimodal (one-peaked) and possibly a little left-skewed, due to the two short bars at left.

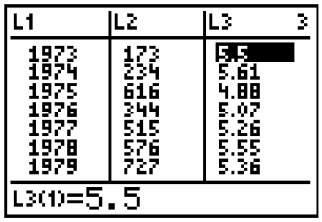
If you are using a TI-89, follow the instructions given above for creating a bar chart. You will need to set the bucket width on the plot definition screen to something reasonable, and modify the WINDOW settings to get a final graph.

Example 1.4 Rock Sole Recruitment: A Time Plot. The next exercise demonstrates how to use a time plot to view data observations that are made over a period of time.

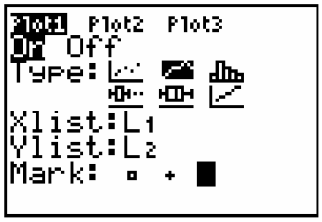
Here are data on the recruitment (in millions) of new fish to the rock sole population in the Bering Sea between 1973 and 2000. Make a time plot of the recruitment.

Year	Recruitment	Year	Recruitment	Year	Recruitment	Year	Recruitment
1973	173	1980	1411	1987	4700	1994	505
1974	234	1981	1431	1988	1702	1995	304
1975	616	1982	1250	1989	1119	1996	425
1976	344	1983	2246	1990	2407	1997	214
1977	515	1984	1793	1991	1049	1998	385
1978	576	1985	1793	1992	505	1999	445
1979	727	1986	2809	1993	998	2000	676

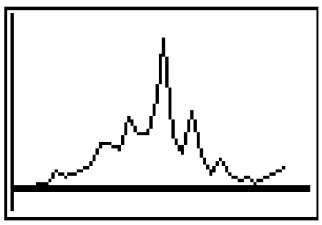
Solution. Enter the data into two lists. For the years, it is easiest to use the seq(function from the LIST menu as described on page 10. Here, I have used lists L1 and L2.



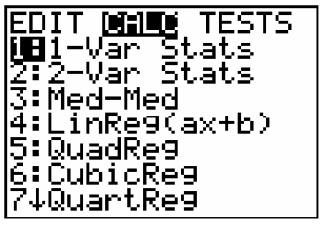
On the STAT PLOTS menu, select Plot1 and define it as at right. This second plot type is a connected scatter plot. If you are using a TI-89, this is the xy-line plot type. The years go on the x-axis of a time plot, so L1 is the Xlist variable, and L2 (the rockfish) is the Ylist. The last data point mark (the single pixel) is recommended for this type of plot, since we really are most interested in seeing the pattern of variation and not the individual points.



Press ZOOM 9 to display the graph. There is no need to adjust the window for this type of plot. We clearly see the rockfish recruitment was small from 1973 but increased fairly steadily until recruitment peaked in 1987. The fish population thereafter seems to have collapsed. A fisheries scientist would be interested in finding out possible reasons to explain this phenomenon.



1.2 Describing Distributions with Numbers

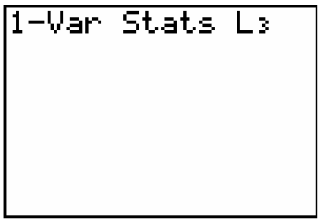


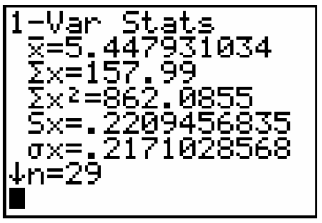
In this section, we will use the 1-Var Stats command from the STAT CALC menu to compute the various statistics of a data set including the mean, standard deviation, and five-number summary. We also will use boxplots and modified boxplots to view these statistics as another way to picture a distribution.

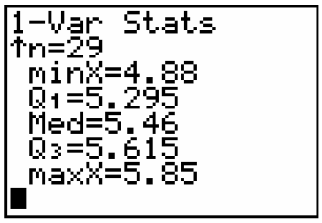
Example 1.5 The Density of the Earth: Finding Summary Statistics. Find $\bar{x}$ and s for Cavendish's data used in Example 1.3 on page 19. Also give the five-number summary and create a boxplot to view the spread.

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

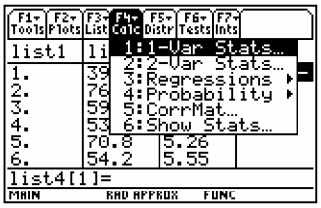
TI-83/84 Solution. Since we still have this data in list L3, we compute the statistics by pressing [STAT] then [▶] to CALC, and finally [ENTER] since option 1:1-Var Stats is highlighted. This transfers the shell of the command to the home screen. We finish the command by telling the calculator which list we are interested in. In this case L3 will be entered as [2nd][3]. *The calculator defaults to calculating statistics for* L1. *It is good practice to always specify the list you are working with.* Press [ENTER] to perform the calculations. The values of $\bar{x}$ and s are then displayed. Scroll down to see the five-number summary.

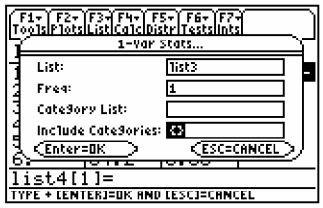


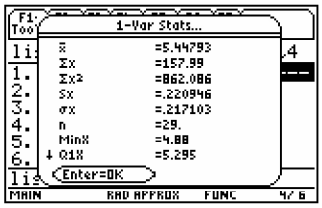




TI-89 Solution. The data have been entered into list3. Press [F4] for the Calc Menu. Option 1:1-Var Stats is highlighted, so press [ENTER]. Locate the list name for which you want Statistics on the [VAR-LINK] screen, highlight the list name and press [ENTER] to move its name onto the definition screen. Finally, press [ENTER] to execute the command. As with the other models, use the down arrow key to complete display of the five-number summary.



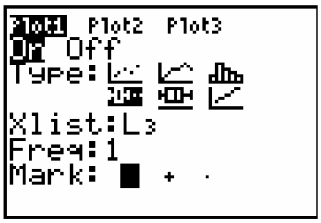


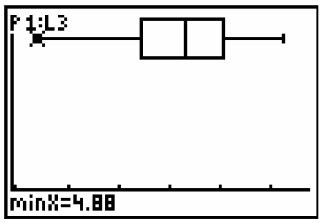


Notice that many of the values have numerous decimal places given. *Your instructor will most likely specify a rounding rule, but the usual rule is to report one more significant digit than was in the original data.* Also, two standard deviation values are given. The first, Sx, is the standard deviation that is denoted by s in the text. Thus, we obtain $\bar{x}$ = 5.448 and s = 0.221 with a five-number summary of 4.88 – 5.295 – 5.46 – 5.615 – 5.85.

Example 1.6 The Density of the Earth: A Boxplot. TI calculators will do two different boxplots: the original that simply uses the five-number summary, and the modified boxplot that uses the 1.5*IQR criteria to identify outliers. We always recommend using the latter, since if a long tail appears, we usually want to know if it's real, or due to outliers. Also, certain distributions that contain outliers may not look like they exist due to scaling.

To make a boxplot of data in a single list, define the plot as below. If you are using a TI-89, use plot type 5:Mod Box Plot. For this type of plot, we do not recommend the single pixel mark for outliers; it is too difficult to see. Select either the box or cross. Press [ZOOM] [9] ([F5] for ZoomData on a TI-89) to display the plot. There is no need to adjust windows. As always, you can use [TRACE] to locate the values on the plot.





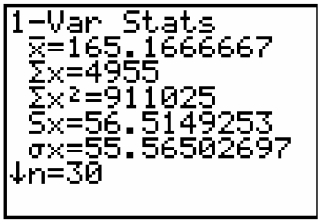
Here we see a longer, left tail to the distribution, indicating more spread of the lowest values. However, since we asked the calculator to identify any outliers, we know there are none as none are indicated.

Example 1.7 Study Time: Side-by-Side Boxplots. The data below give the nightly study time claimed by samples of first-year college men and women. We will first compute the summary statistics for each distribution. We will also make side-by-side boxplots to compare these distributions.

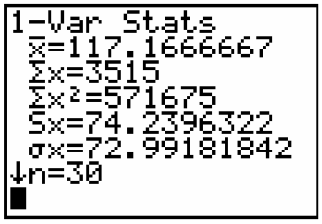
Women					Men				
180	120	180	360	240	90	120	30	90	200
120	180	120	240	170	90	45	30	120	75
150	120	180	180	150	150	120	60	240	300
200	150	180	150	180	240	60	120	60	30
120	60	120	180	180	30	230	120	95	150
90	240	180	115	120	0	200	120	120	180

Solution. We enter the data sets into separate lists and then use the 1-Var Stats command on each list. We have entered the women's data values in to L1 and the men's into L2.

Women:



Men:



The women study longer, on average and have a smaller standard deviation than the men. Using the down arrow, we see that the median for women (175) is also higher than the median for the men (120).

Note: If we have two data sets with an equal number of measurements (like here), then we can compute the statistics of both simultaneously with the 2-Var Stats command from the STAT CALC menu. In this case, enter 2-Var Stats L1,L2. However, this command does not display the five-number summaries.

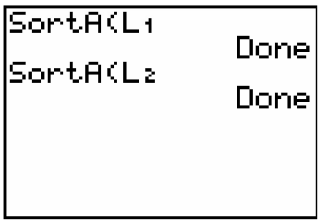
From the five-number summaries, we can compute boundaries (or fences) according to the 1.5**IQR* rule to determine outliers. In each case, we need the values of Q1 and Q3. For the women's study times, Q1 is 120 and Q3 is 180, so these boundaries are

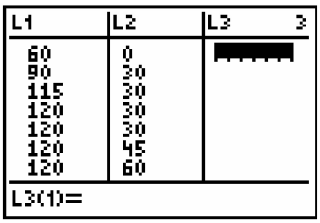
$$120 - 1.5*(180 - 120) = \mathbf{30} \quad \text{and} \quad 180 + 1.5*(180 - 120) = \mathbf{270}$$

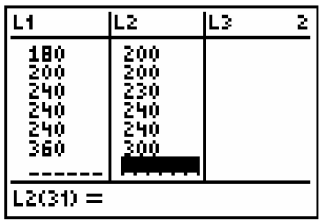
For the men's study times, Q1 is 60 and Q3 is 150, so the fences are:

$$60 - 1.5*(150 - 60) = \mathbf{-75} \quad \text{and} \quad 150 + 1.5*(150 - 60) = \mathbf{285}$$

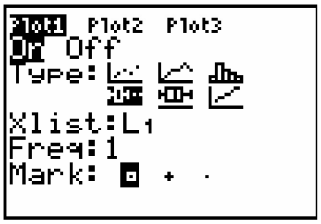
Now we can determine the suspected outliers. For the women, these outliers are any times below 30 minutes or above 270 minutes, while for the men they are any times below –75 minutes or above 285 minutes. To see these values more quickly, we can use the SortA(command from the STAT EDIT menu to sort each list into increasing order. Enter SortA(L1 then SortA(L2. Because these lists have the same size, we can also enter the command SortA(L1,L2. In each case, there are no low outliers, but the time 360 is a high outlier for the women and the time 300 is a high outlier for the men.

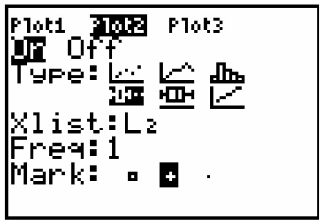


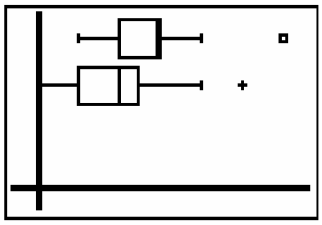




We could have found these outliers much more readily by creating side-by-side boxplots of our two lists of data. This is an exception to the one-plot-at-a-time rule. Below, I define `Plot1` to use the women's data and `Plot2` to use the men's data. The different symbol for any outliers found helps to distinguish which plot is which, but [TRACE] also informs you which list a plot is using. Use the down arrow to move between the plots. Press [ZOOM] [9] to display to graphs. We can clearly see the outlier in both distributions. Notice that Q3 for the women is very close to the median.

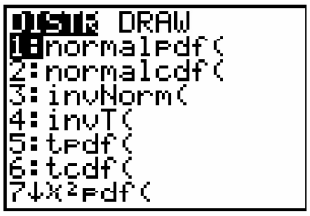






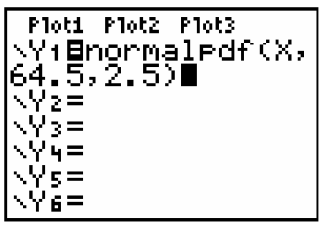
1.3 Density Curves and Normal Distributions

TI calculators have several commands in the `DISTR` menu that can be used for graphing normal distributions, computing normal probabilities, and making inverse normal calculations. In this section, we demonstrate these various functions.

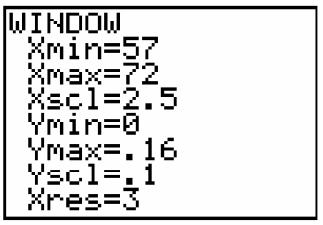


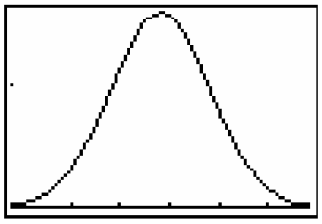
Example 1.8 Women's Heights: Plotting and Shading a Normal Distribution. The distribution of heights of young women are approximately normal with mean $\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches. (a) Plot a density curve for this $N(64.5, 2.5)$ distribution. (b) Shade the region and compute the probability of heights that are within one standard deviation of the mean.

TI-83/84 Solution. (a) We must enter the normal density function `normalpdf(X,`μ`,`σ`)` and adjust the window settings before graphing. Press [Y=] to bring up the function definition screen. Now press [2nd] [VARS] for the `DISTR` menu. Option 1 is `normalpdf`, so press [ENTER] to select it and transfer the shell back to the `Y=` screen. Finish defining the function by entering the mean and standard deviation.



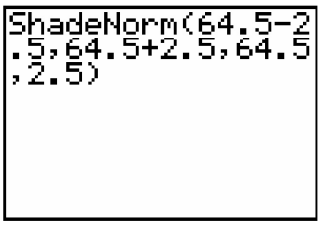
Make certain that all STAT PLOTS are turned off. Press 2nd Y= andmthen 4 to turn all plots off. We will need to size the WINDOW for this plot. Press WINDOW. From the 68-95-99.7 Rule, we know that almost all the area under a Normal curve is within 3 standard deviations of the mean. Three standard deviations here is 3*2.5 = 7.5. We add and subtract that from the mean of 65, so have an Xmin of 57 and Xmax of 72. We have set Xscl to 2.5, so we will see a tick mark for each standard deviation above and below the mean. No probability distribution can be negative (we can't see negative frequencies of something!) so we set Ymin to 0. Setting Ymax for a plot like this is a little harder. The whole area under the curve is 1, so we know Ymax should be less than 1. Play around some to find a good value. Here, I've set Ymax to .16.

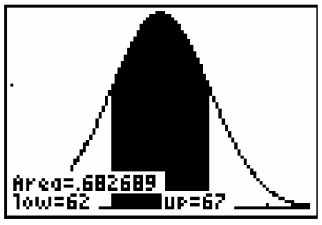




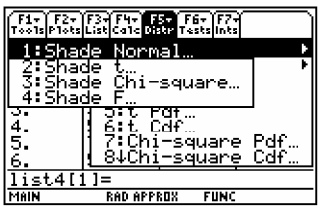
Finally, press GRAPH to display the distribution curve.

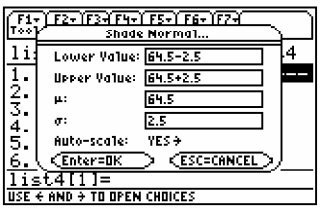
b) Return to the Home screen and press 2nd VARS for the DISTR menu. Press the right arrow key to DRAW. Press ENTER to select option 1:ShadeNorm. You can figure out the heights that are one standard deviation above and below the mean explicitly before entering the command, or you can let the calculator do this for you. Here I'm letting the calculator do that for me. The parameters of the command are low-end, high-end, mean, and sigma. Be sure to separate these with commas. We press ENTER and the desired area will be shaded on the graph created in step a. The calculator also tells us the explicit ends of the area of interest and that 68.27% of women have heights between 62" and 67".

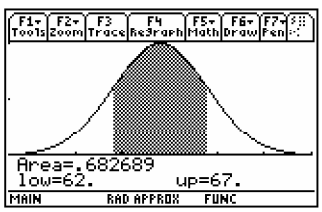




TI-89 Solution. In the Statistics/List Editor, press F5 (Distr). The first option is Shade. Press the right arrow to see more options under this. Press ENTER to select Shade Normal. Enter the lower and upper values of interest (either explicitly, or as I've done here as $\mu-\sigma$ and $\mu+\sigma$) then the values of μ and σ. Select YES to Auto-scale the plot by pressing the right arrow and moving the cursor to highlight YES; this saves having to figure out the window on your own.



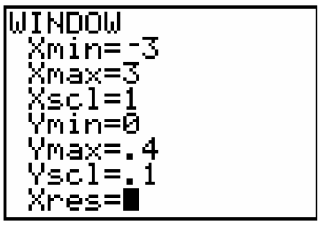


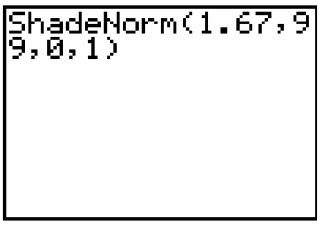


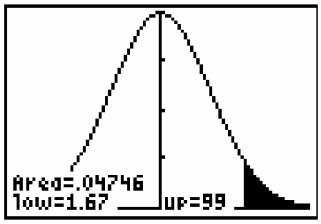
Example 1.9 The Standard Normal Distribution. Let $Z \sim N(0, 1)$ be the standard normal distribution. Shade the areas and find the proportions for the regions (a) $Z > 1.67$, (b) $-2 < Z < 1.67$.

As above, if you are working with a TI-83, you will need to set the window. With a TI-89, this follows immediately from the example above — let the calculator figure out the window for you.

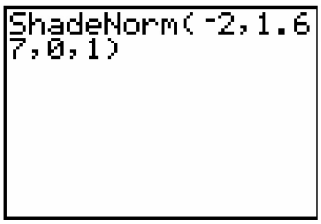
Again, since almost all the area under the curve is within 3 standard deviations of the mean, I have set `Xmin` to –3 and `Xmax` to 3 with `Xscl` 1 (the standard deviation). `Ymin` is 0, and `Ymax` is .4. This window is good for standard Normal distributions. Technically, since the curve extends to ∞ in the positive direction, the high end of interest should be ∞ (`1E99` on the calculator, entered as [1][EE][9][9]), but practically any large number will do. There is almost no area in a normal curve more than 99 standard deviations above the mean.

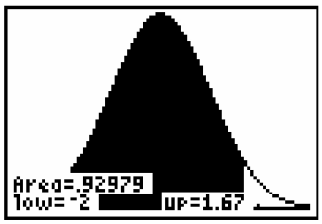






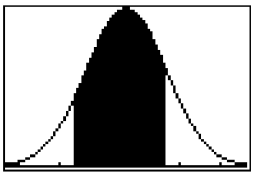
(b) Before drawing a new graph, we will need to clear the shaded part from part a, or we'll just accumulate shading. Press [2nd][PRGM] (`DRAW`) and press [ENTER] to execute option `1:ClrDraw`.





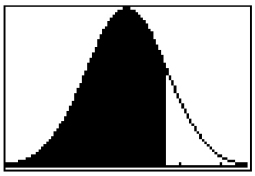
The Normal Distribution and Inverse Normal Commands

For any $N(\mu,\sigma)$ distribution X, we can find probabilities directly with the built-in `normalcdf(` command from the `DISTR` menu. On a TI-89, the command is option 4 on the `Distr` menu. Fill in the boxes as prompted; they will look just like the ones on the `Shade Norm` screen. The command on a TI-83/84 is used as follows.



$P(a < X < b)$

`normalcdf`(a,b,μ,σ)



$P(X < k) = P(X \le k)$

`normalcdf`(–1E 99,k,μ,σ))



$P(X > k) = P(X \ge k)$

`normalcdf`(k,1E 99,μ ,σ))

Example 1.10 More Women's Heights. Find the proportion of American women that are taller than 5'9" (69"). Also find the proportion of women that are shorter than 5' (60 inches).

The screen at right shows the commands and results for these questions. Instead of using 1E99 and –1E99, we could also have used a large positive value and a large negative value with no loss of accuracy. We see that about 3.6% of American women are taller than 5'9"; the same proportion are shorter than 5'.

```
normalcdf(69,1E9
9,64.5,2.5)
      .0359302655
normalcdf(-1E99,
60,64.5,2.5)
      .0359302655
```

Inverse Normal Calculations

To find the value x for which $P(X \le x)$ equals a desired proportion p (an inverse normal calculation), we use the command `invNorm`(p, μ, σ). This is option 3 on the TI-83/84 `DISTR` menu. On a TI-89, press the right arrow to expand the Inverse menu where this is option 1. The following examples demonstrate these commands.

Example 1.11 Women's Heights: Finding Percentiles. How short are the shortest 10% of American women? How tall are the tallest 10% of American women?

Solution: To find the heights of the shortest 10%, we have $P(X \le x) = 0.1$. My first calculation shows that these women are shorter than about 61.3". To find the heights of the tallest 10% of American women, we could use symmetry of Normal distributions (move an equal distance above the mean), but we recognize that $P(X \le x) = 0.9$. The results of my calculation indicate that the tallest 10% are at least 67.7" tall.

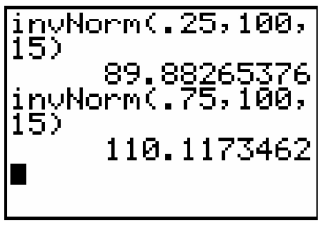
```
invNorm(.1,64.5,
2.5)
      61.29612108
invNorm(.9,64.5,
2.5)
      67.70387892
```

Example 1.12 IQ Scores. The Weschler Adult Intelligence Scale (WAIS) provides IQ scores that are normally distributed with a mean of 100 and a standard deviation of 15. (a) What percent of adults would score 130 or higher? (b) What scores contain the middle 50% of all scores?

Solution. (a) We let $X \sim N(100,15)$ and enter the command at right `normalcdf`(130,99999,100,15). We find that about 2.28% of people should have IQs of at least 130.

```
normalcdf(130,99
999,100,15)
      .022750062
```

(b) If 50% of scores are between a and b, then 25% of scores are below a and 25% of scores are above b. So a is the inverse normal of 0.25 and b is the inverse normal of 0.75. We see that (practically speaking) the middle 50% of people have IQs between 90 and 110.



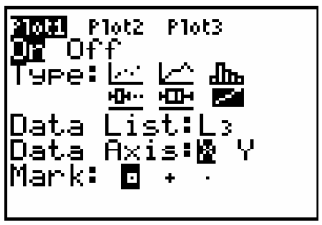
Normal Quantile Plots

Normal Quantile plots on TI-calculators plot the data values against the z-score that value would have if the data came from a normal distribution. This used to be an extremely tedious plot to create by hand, but technology makes these easy. If the data do come from an (approximately) normal distribution, the plot of points should be a straight line.

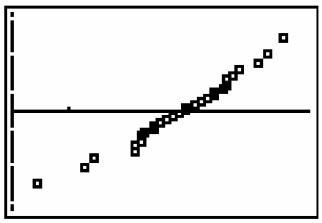
Example 1.13 The Density of the Earth: A Normal Plot. Make a normal quantile plot of Cavendish's data that were used in Examples 1.3 and 1.5.

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

TI-83/84 Solution. We still have these data in list L3. Normal quantile plots are the last plot type on a plot definition screen. You have the option of either the x- or y-axis containing the data, which is a matter of personal preference. Since we hope to see a straight line result, it really makes no difference, but this author is accustomed to data on the x-axis, so we have selected that. You (as always) have the option of selecting the mark for each data point.

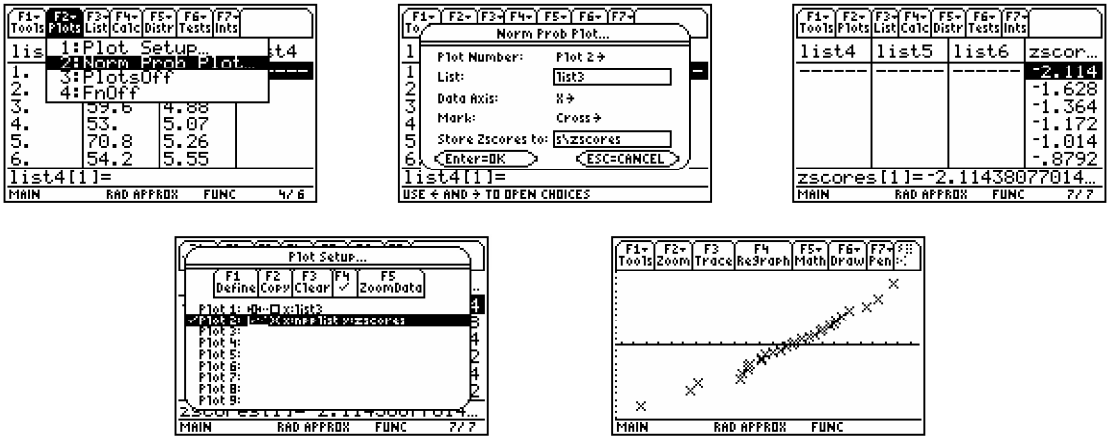


Press ZOOM 9 to display the graph. Recall from the boxplot of these data (page 22) that there was a long left tail. In this plot, these data values seem to split away somewhat from the bulk of the data. While they weren't outliers, they are a little too far out for this data to be perfectly normal. We might be happy to call it approximately normal.



TI-89 Solution. On TI-89 calculators there is an intermediate step in creating a normal quantile plot that makes the z-score computations more explicit. Since we still have Cavendish's data in list3, from the Statistics/List Editor application press F2 for the Plots menu. Select option 2:Norm Prob Plot. You can select the plot number (it defaults to one higher than what is already defined), the list to use (use [VAR-LINK] to insert

the list name), the axis that will contain the data, and a list to contain the z-scores (we recommend taking the defaults). Press ENTER to perform the calculations. Z-scores for each data value are stored and displayed on the editor screen. Return to the Plot Setup menu. Uncheck using F4 or clear any other plots. Move the cursor to highlight the normal plot, then press F5 to display the final plot.



Example 1.14 A Normal Quantile Plot for a Uniform Distribution. Certain types of distributions have classical normal quantile plots. Generate 100 observations from a Uniform[0,1] distribution and display its normal plot.

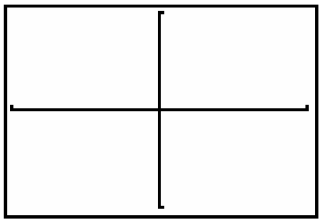
Solution. Press MATH. Arrow to PRB. The first option on this menu generates uniform random numbers on [0,1]. Since we want 100 of them, we will use the command rand(100)STO▸L4.

The calculator will echo back the first one or two of these. To look at them further, use the editor. Define the normal plot as before to use this list. Press ZOOM 9 to display the plot. Here we see the classical bends of a uniform distribution at each end of the plot. This is because uniform random variables have abrupt ends — there is no gradual tapering of the distribution as there is with a normal distribution.

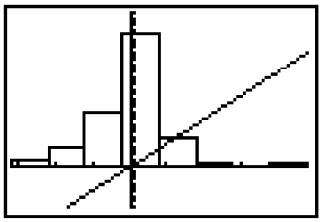


1.4 Common Errors

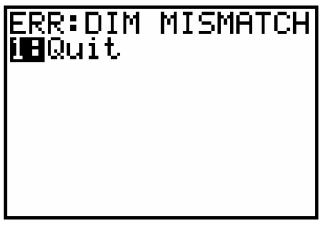
There's no picture!
Seeing something like this (or a blank screen) is an indication of a windowing problem. This is usually caused by pressing [GRAPH] using an old setting. Try pressing [ZOOM][9] to display the graph with the current data. This error can also be due to having failed to turn the plot "On."



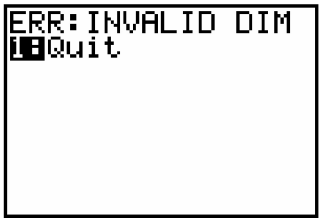
What's that weird line (or curve)?
There was a function entered on the [Y=] screen. The calculator graphs everything it possibly can at once. To eliminate the line, press [Y=]. For each function on the screen, move the cursor to the function and press [CLEAR] to erase it. Then redraw the desired graph by pressing [GRAPH].



Err: Dim mismatch
This common error results from having two lists of unequal length. Here, it pertains either to a histogram with frequencies specified or a time plot. Press [ENTER] to clear the message, then return to the statistics editor and fix the problem.



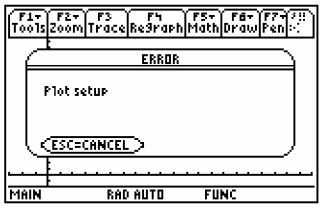
Err: Invalid dim
This problem is generally caused by reference to an empty list. Check the statistics editor for the lists you intended to use, then go back to the plot definition screen and correct them.



Err: Stat
This error is caused by having two stat plots turned on at the same time. What happened is the calculator tried to graph both, but the scalings are incompatible. Go to the STAT PLOT menu and turn off any undesired plot.

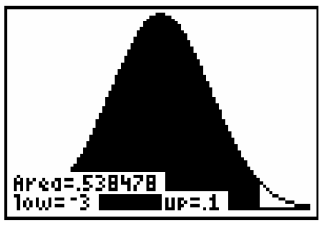


Plot setup
This is the TI-89 equivalent of the STAT error above. It is caused by having two stat plots turned on at the same time. The calculator tried to graph both plots, but the scalings are incompatible. Go to the Stat plots menu and turn off any undesired plot by moving the cursor to that plot, and pressing [F4].



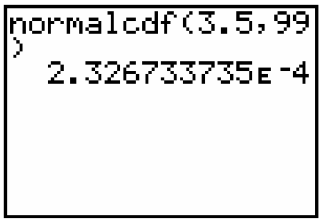
Why is my curve all black?

For the standard normal curve, the graph indicates that well over half of the area is of interest between –3 and .1. The message at the bottom says the area is 53.8%. This is a result of having failed to clear the drawing between commands. Press 2nd PRGM then ENTER to clear the drawing, then reexecute the command.



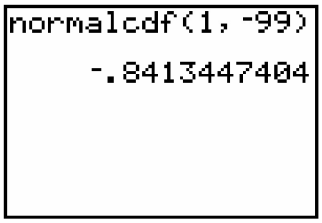
Probability more than 1

This is not possible. If the results look like the probability is more than one, check the right side of the result for an exponent. Here it is –4. That means the leading 2 is really in the fourth decimal place, so the probability is 0.0002. The chance of a variable being more than 3.5 standard deviations above the mean is about 0.02%.



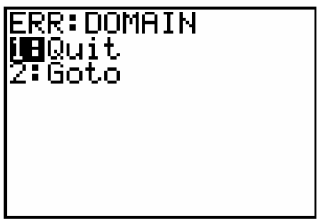
Negative probability

This is not possible. The low and high ends of the area of interest have been entered in the wrong order. As the calculator does a numerical integration to find the answer, it doesn't care, but you should.



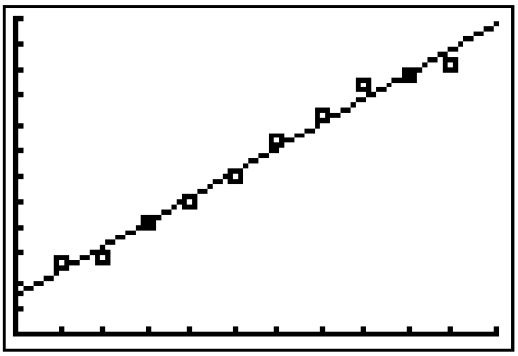
Err: Domain

This message comes as a result of having entered the `invNorm` command with parameter 90. (You wanted to find the value that puts you into the top 10% of women's heights, so 90% of the area is to the left of the desired value.) The percentage must be entered as a decimal number. Reenter the command with parameter .90.



CHAPTER

2



Looking at Data—Exploring Relationships

2.1	Scatterplots
2.2	Correlation
2.3	Least-Squares Regression
2.4	Cautions About Correlation and Regression
2.5	Common Errors

Introduction

In this chapter, we use TI calculators to graph the relationship between two quantitative variables using a scatterplot. We then show how to compute the correlation and find the least-squares regression line through the data. Last, we show how to work with the residuals of the regression line.

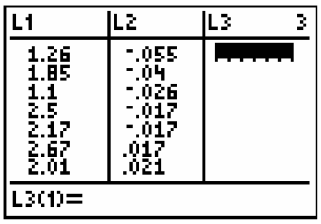
2.1 Scatterplots

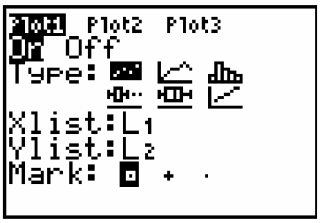
We begin by showing how to make a scatterplot of two quantitative variables along the x and y axes so that we may observe if there is any noticeable relationship. In particular, we look for the strength of the linear relationship.

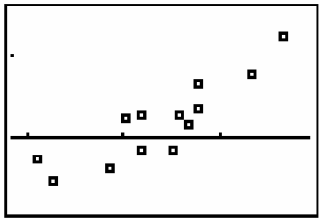
Example 2.1 Brain Activity and Stress. It has been suggested that emotional stress leads to increased brain activity. Make a scatterplot of brain activity level against social distress score.

Subject	Social distress	Brain activity	Subject	Social distress	Brain activity
1	1.26	–0.055	8	2.18	0.025
2	1.85	–0.040	9	2.58	0.027
3	1.10	–0.026	10	2.75	0.033
4	2.50	–0.017	11	2.75	0.064
5	2.17	–0.017	12	3.33	0.077
6	2.67	0.017	13	3.65	0.124
7	2.01	0.021			

TI-83/84 Solution. We first enter the data into the STAT EDIT screen. Here we use L1 for the social distress scores, plotted on the x axis, and use L2 for the brain activity levels that will be plotted on the y axis. Define PLOT1 on the STAT PLOTS menu to use the first (scatterplot) plot type as below. You have your choice of marks for each data point, but we do not recommend the last (single pixel) as it is hard to see. Press [ZOOM][9] to display the plot.

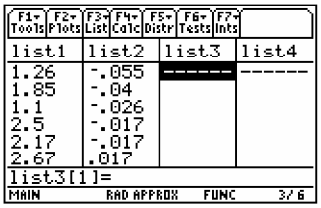


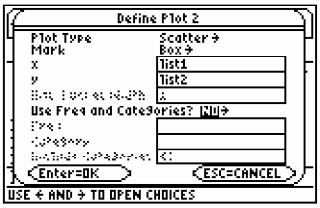


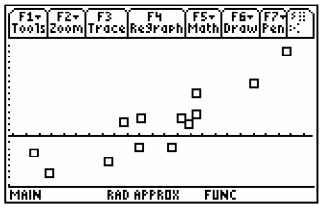


TI-89 Solution. With a TI-89, the sequence required to generate the scatterplot is very similar . We have entered the distress scores into list1 and the brain activity levels into list2. Press [F2] for Plots and select option 1:Plot Setup. Clear or uncheck ([F4]) any unneeded plots. Select a plot to define and press [F1] to proceed. The plot type is Scatter. You can use the right arrow key to expand the plot type options, if needed. Select a plot symbol for each data point in a like manner. Use [2nd][–] ([VAR-LINK]) to access the lists of variable names and select list1 for the x variable and list2 for the y variable. We are not using separate lists of frequencies or categories, so this should be

set to NO. Press [ENTER] to complete the definition and return to the Plot Setup screen. Press [F5] to display the plot.





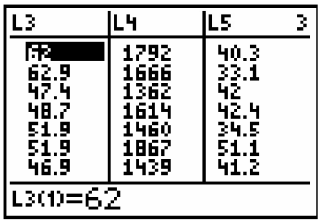


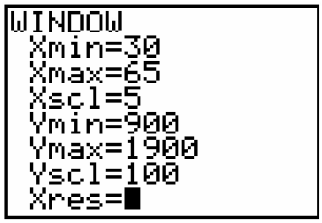
We see that as the social distress score increases, the brain activity level generally tends to increase also.

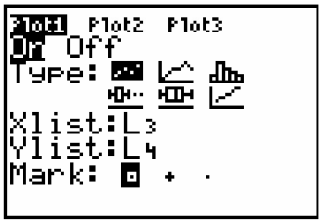
Example 2.2 Body Mass and Gender. We would like to examine whether or not there seems to be a difference in metabolic rate versus body mass for males and females. We will use two scatterplots with different symbols for males and females, and display both at once.

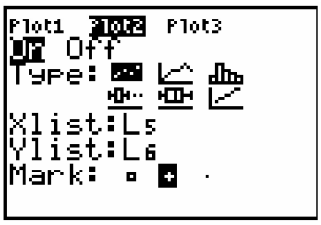
Sex	Mass	Rate	Sex	Mass	Rate
M	62.0	1792	F	40.3	1189
M	62.9	1666	F	33.1	913
M	47.4	1362	F	42.0	1418
M	48.7	1614	F	42.4	1124
M	51.9	1460	F	34.5	1052
M	51.9	1867	F	51.1	1347
M	46.9	1439	F	41.2	1204
			F	54.6	1425
			F	50.6	1502
			F	36.1	995
			F	42.0	1256
			F	48.5	1396

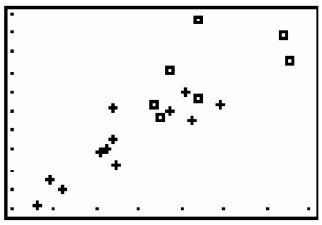
Solution. We first enter the mass and rate of just the males into lists L3 and L4 respectively. Then we enter the mass and rate of the females into lists L5 and L6. Here, we will need to adjust the WINDOW so that the X range includes all the masses and the Y range includes all the rates. We adjust the STAT PLOT settings for Plot1 to obtain the scatterplot of L3 versus L4 (the males), and adjust the STAT PLOT settings for Plot2 to obtain the scatterplot of L5 versus L6 (the females). Since we have explicitly sized the window, press [GRAPH] to display the plots.









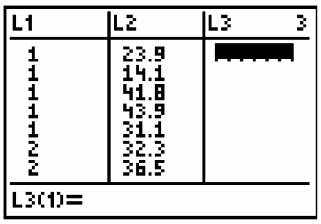


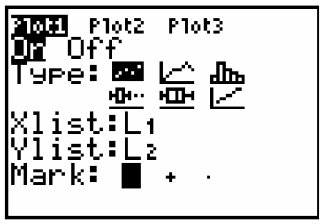
We see a linear pattern between the two genders. This is clearly stronger for women who have both lower body mass and metabolism. The men's data at the upper right is much more scattered (a weaker relationship).

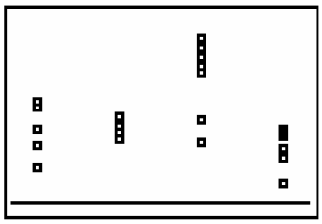
Example 2.3 Sector Fund Returns. How does the return on investment vary among sector mutual funds? Data in the table below are annual total returns from several sector funds. We will make a plot of the total return against market sector. We'll also compute the mean return for each sector, add the means to the plot, and connect the means with line segments, so the averages will stand out more

Market sector	Fund returns (percent)						
Consumer	23.9	14.1	41.8	43.9	31.1		
Financial services	32.3	36.5	30.6	36.9	27.5		
Technology	26.1	62.7	68.1	71.9	57.0	35.0	59.4
Natural resources	22.9	7.6	32.1	28.7	29.5	19.1	

Solution. We will plot the market sectors on the x axis as the values 1, 2, 3, and 4. Because there are multiple returns for each sector, we enter each of the values 1 through 4 as many times into list L1 as there are returns for that sector. We enter the corresponding returns into list L2. We define the scatterplot as we have done previously, and press ZOOM 9 to display the plot (there is no need to specially size this one).



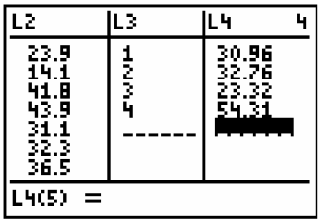


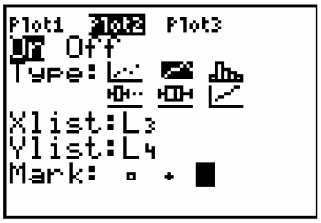


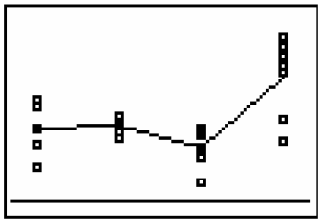
So far, we see that the returns in sector 3 (Technology) are much more variable than the others there is potential for both greater and smaller returns. Natural resources seems to have the lowest returns, and Financial Services the most consistent (least variability).

Data for each sector was reentered into a new list and the mean calculated for each. We have now entered the category numbers and means. We will define a new connected scatterplot to use the data in lists L3 and L4, this scatterplot will use the single pixel mark, since we don't want to have the appearance of an extra data point. Since the graph

window is already sized for these data, pressing either GRAPH or ZOOM 9 will display the plot.







2.2 Correlation

In this section, we compute the correlation coefficient r between paired quantitative variables. For all TI calculator models, to obtain correlations, the linear regression must be computed

If you are using a TI-83 or -84, we must make sure that the calculator's diagnostics are turned on. Enter the CATALOG (2nd 0) and press x^{-1} to advance to functions beginning with the letter D. Scroll down until you find the DiagnosticOn command. Press ENTER to bring the command to the Home screen, then press ENTER again to execute the command.

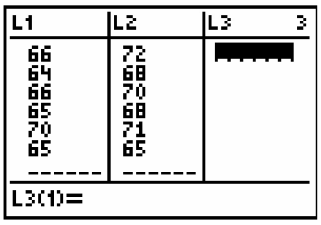


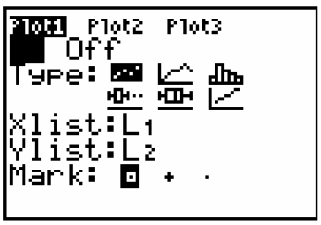


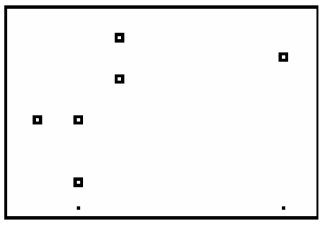
Example 2.4 Dates' Heights. The table below gives the heights in inches for a sample of women and the last men whom they dated. (a) Make a scatterplot. (b) Compute the correlation coefficient r between the heights of these men and women. (c) How would r change if all the men were 6 inches shorter than the heights given in the table?

Women (x)	66	64	66	65	70	65
Men (y)	72	68	70	68	71	65

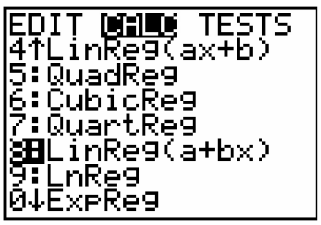
TI-83/84 Solution. (a) We enter the heights of the women into list L1 and the heights of the men into list L2, adjust the STAT PLOT settings, and display the graph using ZOOM 9. If you have just completed working through the previous example, be sure to turn Plot2 OFF.

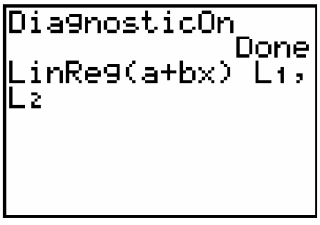


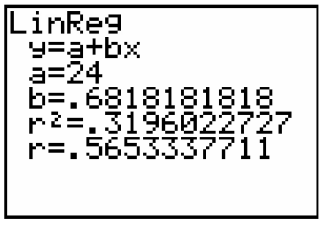




(b) To compute the correlation, we use the `LinReg(a+bx)` command (option 8) from the `STAT CALC` menu. Enter the command `LinReg(a+bx) L1,L2`. The `LinReg(ax+b)` command (item 4) will also compute the correlation. We are actually computing the linear regression here (but won't examine it yet) in order to get the correlation.

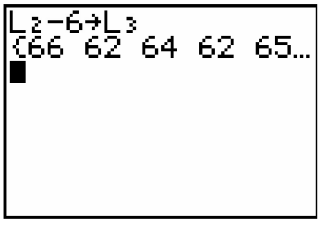


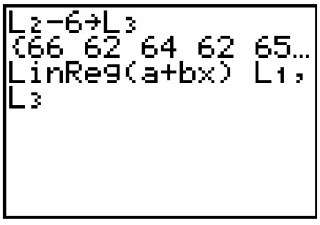


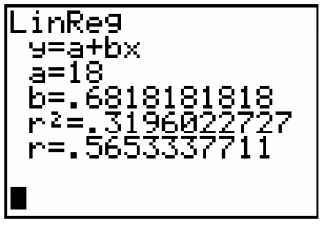


The correlation between the men's and women's heights is $r = 0.565$.

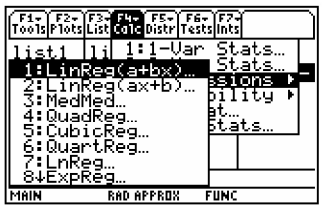
(c) In order to get heights for all the males that are 6 inches shorter than they originally were, we use the calculator to subtract 6 from each entry in `L2` and store the result into `L3`. The `STO▸` key is above `ON` at the lower left of the keyboard. After pressing `ENTER` to perform the calculation, the first few values of the new list are shown. To view the complete list, use the editor. Then we use the command `LinReg(a+bx) L1,L3` to see that r has not changed.

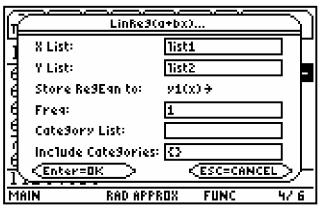


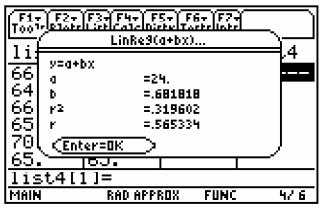




TI-89 Solution: The data have been entered into `list1` and `list2`. Press `F4` for the `Calc` menu and select option `3:Regressions`. The submenu of all possible regressions will open. Select option `1:LinReg (a+bx)`. In the dialog box, we have given our list names, and also asked that the regression equation be stored in the `y1(x)` function. Press `ENTER` for the results.







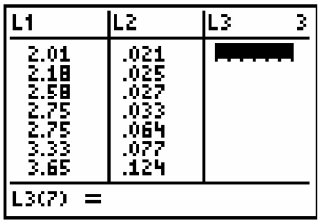
2.3 Least-Squares Regression

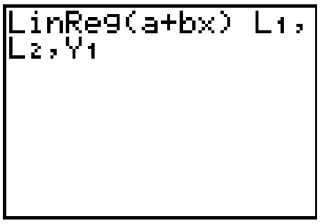
In this section, we will compute the least-squares line of two quantitative variables and graph it through the scatterplot of the variables. We will also use the line to predict the *y*-value that should occur for a given *x*-value.

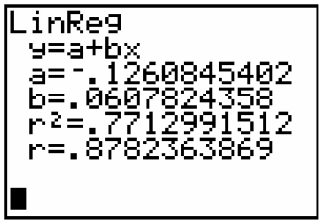
Example 2.5 More Brain Activity and Stress. The data from Example 2.1 are repeated below. (a) What is the equation of the least-squares regression line for predicting brain activity from a social distress score? Make a scatterplot with this line drawn on it. (b) Use the equation of the regression line to get the predicted brain activity level for a distress score of 2. (c) What percent of the variation in brain activity among these subjects is explained by the straight-line relationship with social distress score?

Subject	Social distress	Brain activity	Subject	Social distress	Brain activity
1	1.26	–0.055	8	2.18	0.025
2	1.85	–0.040	9	2.58	0.027
3	1.10	–0.026	10	2.75	0.033
4	2.50	–0.017	11	2.75	0.064
5	2.17	–0.017	12	3.33	0.077
6	2.67	0.017	13	3.65	0.124
7	2.01	0.021			

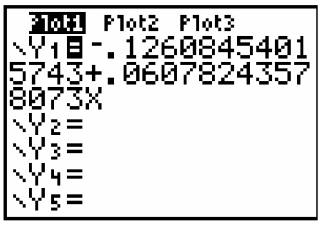
Solution. (a) We obtain the linear regression line using the same `LinReg(a+bx)` command that computes the correlation. After entering data into lists, say `L1` and `L2`, we would enter the command `LinReg(a+bx) L1,L2`. However, in order to store the equation and be able to use it for graphing or prediction, we will add a new parameter to the LinReg command. Add a comma after `L2` and press [VARS], press [▸] to move the highlight to Y-VARS, press [ENTER] to select option 1:Function and [ENTER] to select `Y1`. Finally, press [ENTER] to perform the calculations.

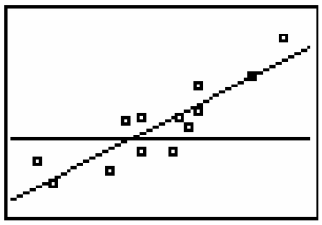




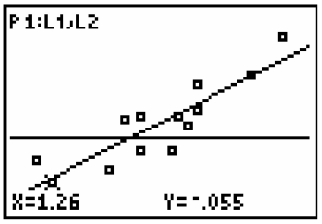


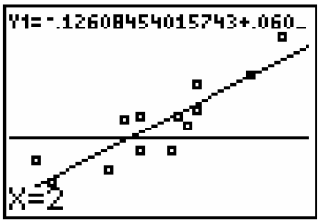
You can see the equation of the line stored by pressing [Y=]. Our equation is Brain-Activity = –0.126+0.061*Distress. To graph the data along with the regression line, press [ZOOM] [9] ([F5] for ZoomData from the Plot Setup Screen on a TI-89).

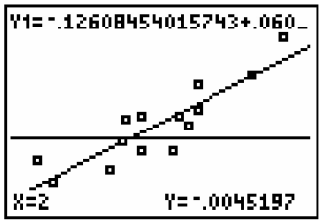




(b) With the equation of the regression line computed and displayed on the graph as above, we can use the [TRACE] ([F3] on a TI-89) to evaluate the function for a specific x value. Press [TRACE]. At the upper right, the active plot is displayed, which indicates we are tracing the scatterplot. Press [▼] to switch to the line. The top line will change to display the equation of the line. Type in 2 (the distress value we want to find a brain activity for) and press [ENTER]. We see that a distress score of 2 gives a predicted brain activity of –0.0045.



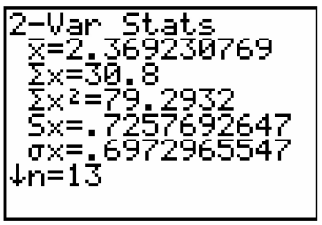


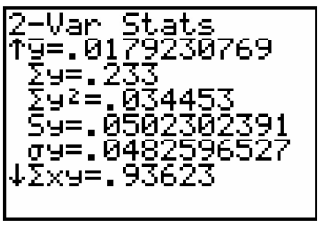


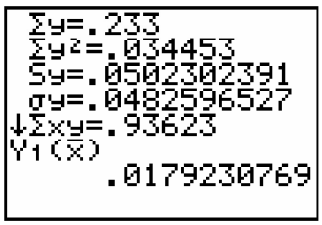
Alternately, we can access the Y1 function from the Home screen. To do so, press [VARS], arrow right to Y-VARS, enter [1] for Function, enter [1] for Y1, then enter the command Y1(2). Press [ENTER] to perform the calculation.



We can also verify that the point $(\bar{x}, \bar{y})$ is on the regression line; but first we must compute the statistics. We can do so simultaneously with the 2-Var Stats command from the STAT CALC menu, because the two data sets are the same size. Enter the command 2-Var Stats L1,L2. Then enter Y1($\bar{x}$) by recalling $\bar{x}$ from the [VARS] Statistics menu.







(c) With the calculator's diagnostics turned on, the `LinReg(a+bx)` command also displays the values of r and r^2. In this case, $r^2 = 0.7713$. Thus, 77.13% of the variation in brain activity among these subjects is explained by the straight-line relationship with social distress score.

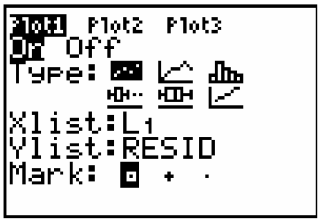
2.4 Cautions About Correlation and Regression

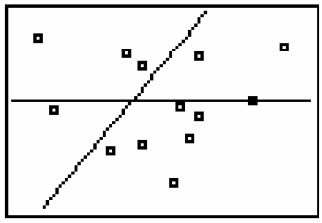
We now complete an exercise to demonstrate how to work with the residuals of a least-squares regression line. These are computed for each regression and stored in a list called `RESID`. Ideally, these will be randomly distributed around the x-axis ($y = 0$ line). Any indications of pattern (curvature, widening (or narrowing) as x-values increase) show a violation of assumptions for the regression. They can also be used to help identify outliers and potentially influential points in a regression.

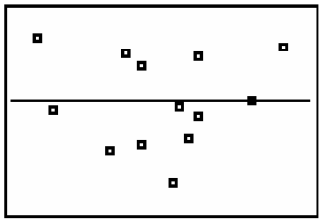
Example 2.6 More Brain Activity and Stress. We continue with our example on distress and brain activity. We want to obtain the residuals scatterplot against our x (predictor) variable.

Solution. We first define the scatterplot as shown below. To access the list named RESID, press [2nd][STAT], and locate the list name. Press [ENTER] to transfer the list name to the plot definition screen. Press [ZOOM][9] to display the plot. Oops! The regression line shown on our plot below is *not* part of the residuals plot! Go to the [Y=] screen and press [CLEAR] to erase the equation. Redisplay the plot using either [ZOOM][9] or [GRAPH].

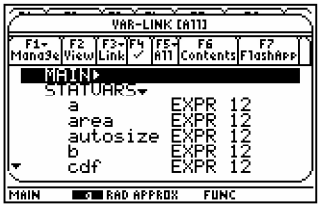
This plot is pretty much ideal. Points are relatively evenly spread around the $y = 0$ axis and seem randomly distributed.

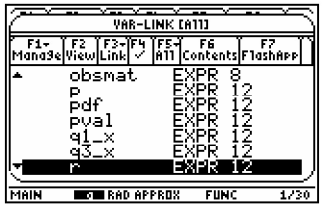


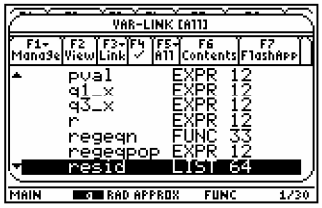




TI-89 Note: To locate the `resid` list on the [VAR-LINK] screen, press ◀ to collapse the `MAIN` folder. Press the down arrow to get inside the `STATVARS` folder, then press [2] (R in alpha mode) to get to the portion of these variables that begin with the letter r. Move the arrow down until you find the list and press [ENTER] to select it and transfer this name to the plot definition.



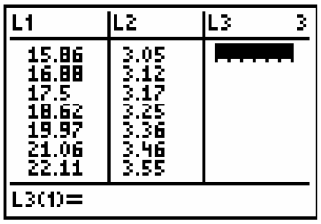


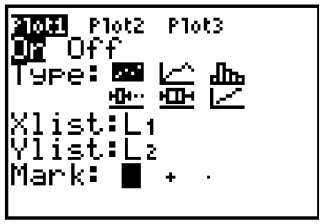


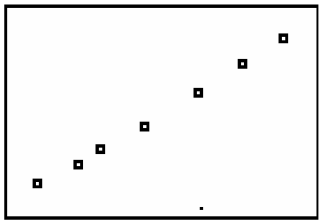
Example 2.7 Runners' Stride Rates. The following table gives the speeds (in feet per second) and the mean stride rates for some of the best female American runners. We'll perform a complete analysis of these data, attempting to predict stride rate (y) using the speed (x) of the runner.

Speed	15.86	16.88	17.50	18.62	19.97	21.06	22.11
Stride rate	3.05	3.12	3.17	3.25	3.36	3.46	3.55

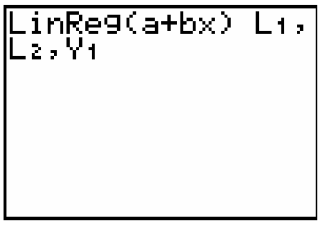
Solution. The data have been entered into L1 and L2. The initial data plot is defined and displaycd using ZOOM 9.

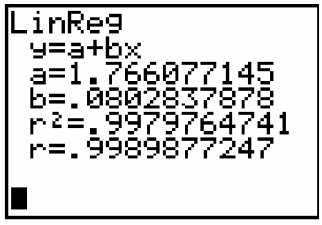


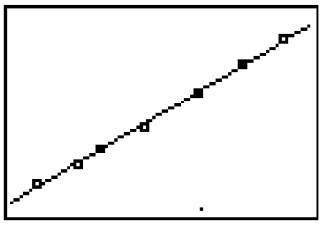




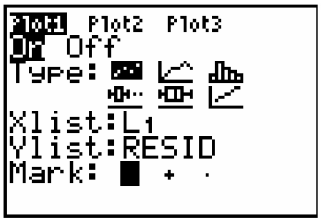
The plot appears to be very linear. We'll calculate the regression, storing the equation into Y1, and display the line on the graph.







Our regression equation is Stride-rate = 1.766 + 0.080*Speed. This appears to be an extremely strong relationship, since the linear regression explains 99.9% (r^2) of the variation in stride rate. We're not done, however. Examining the residuals plot should always be the next step — is our model the "correct" one for these data? The residuals plot is defined and displayed using ZOOM 9.

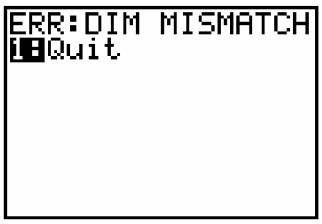




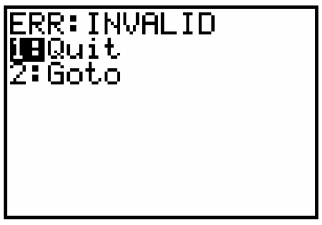
These residuals indicate a clear curve. A straight line model is *not* the correct one!

2.5 Common Errors

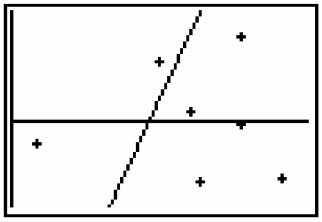
Err:Dim mismatch
We've seen this one before. Press [ENTER] to quit. This error means the two lists referenced (either in a plot or a regression command) are not the same length. Go to the STAT editor and fix the problem.



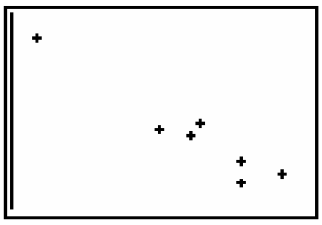
Err: Invalid
This error is caused by referencing the function for the line when it has not been stored. Recalculate the regression being sure to store the equation into a *y* function.



What's that weird line?
This error can come either in a data plot (an old line still resides in the [Y=] screen) or the stored regression line is showing in the residuals plot, as shown here. As stated earlier, the regression line is *not* part of the residuals plot and shows only because the calculator tries to graph everything it knows about. Press [Y=] followed by [CLEAR] to erase the unwanted equations, then redraw the graph by pressing [GRAPH].

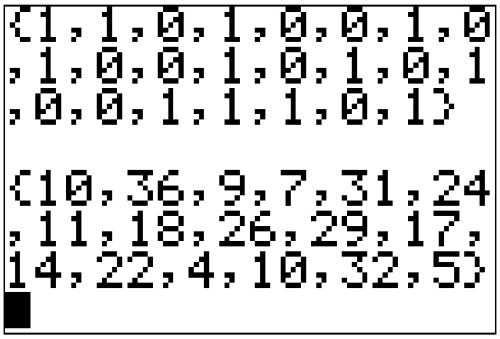


This doesn't look like a residuals plot!
Residuals plots *must* be centered around $y = 0$. This error is usually caused by confusing which list contains the *y*'s and which the *x*'s in computing the regression or in defining the residuals plot. Go back and check which list is which, and recompute the regression, or redefine the plot.



CHAPTER

3



Producing Data

3.1	First Steps
3.2	Design of Experiments
3.3	Sampling Design
3.4	Toward Statistical Inference

Introduction

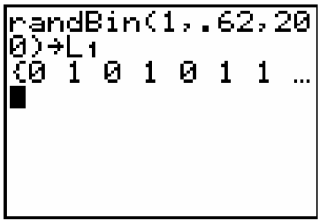
In this chapter, we use TI calculators to simulate the collection of random samples. We also provide a supplementary program that can be used to draw a random sample without repeats from a list of integers numbered from *m* to *n*. Good data collection practice involves randomly selecting individuals from the population, or randomly assigning treatments in a controlled experiment. The randomization can be done with a random digits table, a calculator, or a computer. When your text says "start on line xx of table b" the sample drawn in that manner will *not* be random — this is merely a mechanism to be able to write an answer for the back of the book.

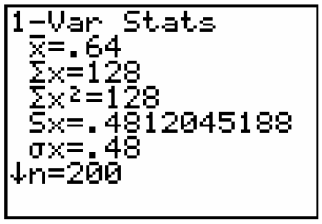
3.1 First Steps

In this section, we demonstrate how to generate count data, or *Bernoulli trials*, for a specified proportion p. The data simulates observational "Yes/No" outcomes obtained from a random survey. To generate the data, we use the `randBin` command from the `MATH PRB` menu (option 7).

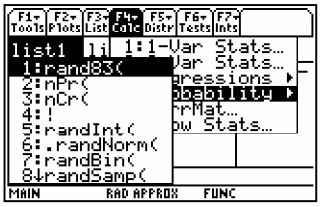
Example 3.1 Simulating a Survey. Suppose that 62% of students hold a part-time or full-time job at a particular university. Simulate the results of a random survey of 200 students and determine the sample proportion of those who have a job.

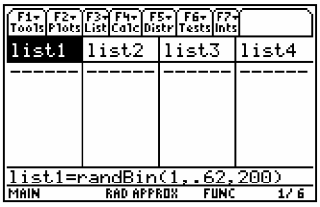
TI-83/84 Solution. To generate a random list of 1 and 0 responses ("Yes/No"), enter the command `randBin(1,p,n)` STO▶ `L1`, where p is the specified proportion of yeses (these will be 1s in our data), and n is the desired sample size. Here, use `randBin(1,.62,200)` STO▶ `L1`. Then enter the command `1-Var Stats L1`. Notice after the command is executed, the first few values are shown on the screen. To be able to see the whole list, use the Statistics Editor. In our sample, we actually had 64% "successes."

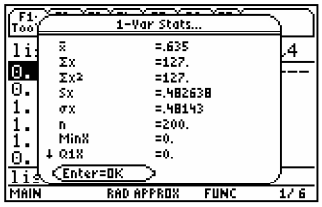




TI-89 Solution. In the `Stats/List Editor` application, move the cursor to highlight the list name that will hold our random data. Press F4 for the `Calc` menu, arrow down to option `4:Probability`, and press the right arrow to expand the options. Press 7 to select randBin. Complete the command by entering the parameters 1,.62,200 then close the parentheses and press ENTER to execute. You will see the first few entries in the generated list, and can page down to view the entire list, if desired. We show the results of performing 1-Var Stats on this generated data to show that results are random — here, we had 63.5% 1s.



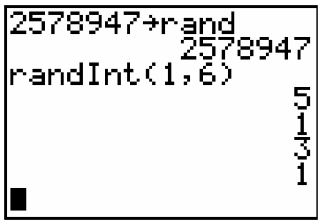




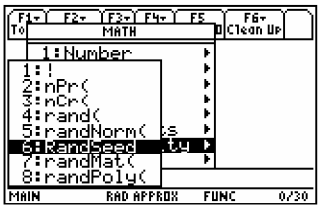
Using the calculator to generate random "samples" is not truly random. These are really *pseudo*random numbers. The calculator uses a value called a "seed" to control the sequence. This seed is changed each time a random number is generated, so results should appear random each time.

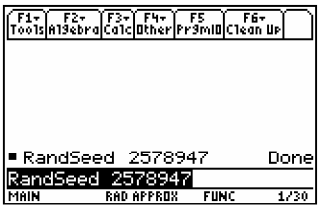
Example 3.2 Changing the Seed. We really want random numbers each time. However, just as your text will instruct you to "start in line xx" when using the table of random digits, your instructor may ask that you use a particular seed so that each student will get the same random numbers. Good seed numbers are large, odd, and preferably prime numbers. The method of setting the seed varies with calculator type. We'll follow the seed setting with a `randInt` command (also found on the `MATH PRB` menu) to simulate throwing a die.

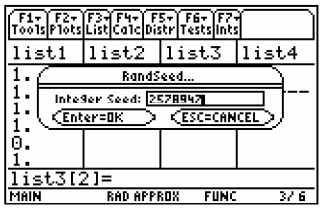
TI-83/84 Solution. Type the desired seed number on the home screen. Press [STO▸]. Then, from the `MATH PRB` menu, select option `1:rand`. Press [ENTER] and the calculator will echo the seed number back. Immediately follow this with the `randInt(1,6)` to generate numbers randomly from 1 to 6 (inclusive). To keep generating values, keep pressing [ENTER]. If you use my seed as at right, you will get the same "random" numbers.



TI-89 Solution. The seed can be set from the home screen using option `6:randseed` from the `Math Probability` menu. Type in the desired seed and press [ENTER] to see the `Done` message. This can also be done in the Stats/List Editor application using option `A:randseed` from the `Calc Probability` menu.

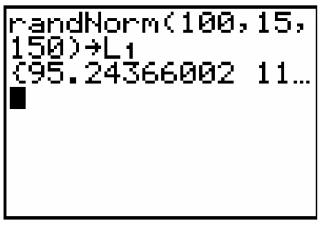


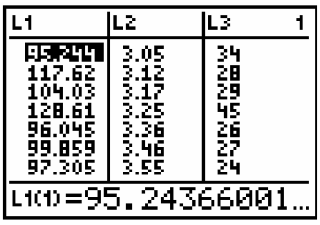


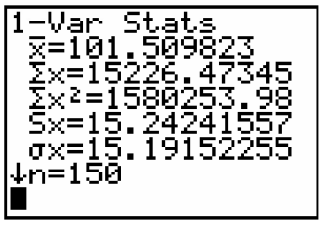


Example 3.3 Simulating IQ Scores. Generate 150 observations from a $N(100,15)$ distribution. This distribution will mimic scores for individuals on the Wechsler Adult Intelligence Scale. Compute the sample statistics to compare $\bar{x}$ with 100 and to compare s with 15.

Solution. From the `MATH PRB` menu, select `6:randNorm(`. Complete the command by entering the parameters `100,15,150)`[STO▸]`L1` and then compute the sample statistics. If you are using a TI-89, the procedure is essentially the same as used above — select option `6:.randnorm` from the `Calc Probability` menu. Notice that on the home screen, many more decimal places are shown than in the list editor. One point to make here, is that normal random variables are truly continuous (many decimal places are possible) while the IQ scores are really discrete (or at least rounded).







My sample mean (101.5) and standard deviation (15.24) are close to the parameter values, but not exactly the same.

3.2 Design of Experiments

Assigning treatments for a randomized experiment (or selecting individuals for a sample) can easily be done using the `randInt` command we just used. One small drawback is that "selected" individuals can be selected more than once.

Below, we provide a TI-83/84 supplementary program that can be used to choose subjects at random from an enumerated group without any repeated numbers. The TI-89 has a built-in function `randSamp` on the Calc Probability menu that samples with replacement from a list of numbers.

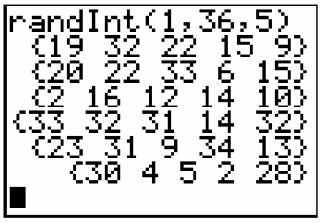
The RANDOM Program

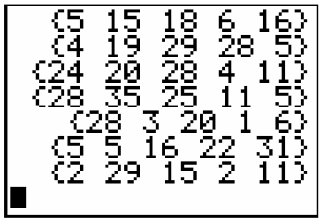
```
PROGRAM:RANDOM
:Disp "LOWER BOUND"
:Input M
:Disp "UPPER BOUND"
:Input N
:Disp " HOW MANY?"
:Input R
:ClrList L4
:seq(J,J,M,N)→L1
:For(I,1,R)
:ClrList L5
:randInt(1,N-M+2-I)→A
:L1(A)→L4(I)
:1→K
:While K<A
:L1(K)→L5(K)
:1+K→K
:End
:A→K
:While K≤N-M+1-I
:L1(K+1)→L5(K)
:1+K→K
:End
:L5→L1
:End
:L4→L1
:ClrList L5,L4
:ClrHome
:Output(1,2,L1)
:Stop
```

The `RANDOM` program can be used to choose a random subset of k subjects from a group that has been numbered from m to n. It also can be used to permute an entire set of n subjects so that the group can be assigned randomly to blocks. The program displays the random choices and also stores the values into list `L1`.

Example 3.4 Random Treatment Assignment. We have 36 subjects numbered 1 through 36 available for a small clinical trial. We want to assign them randomly to four treatment groups, each of size 9.

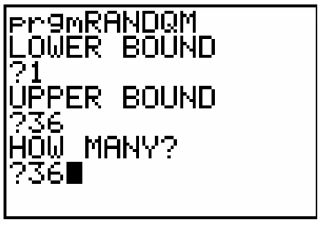
Basic Solution. Use the randInt(function from the MATH PRB menu to assign the subjects. The first nine numbers generated will receive the first treatment; the second nine the second treatment, and the third nine will get the third treatment. At that point, anyone not already assigned will get the fourth treatment. We could generate a long list of numbers into a list as has been done previously (this is preferred if using a TI-89), or simply keep hitting ENTER for numbers singly, but for ease of viewing, we have used the command randInt(1,36,5) which will give 5 numbers each time we press ENTER.

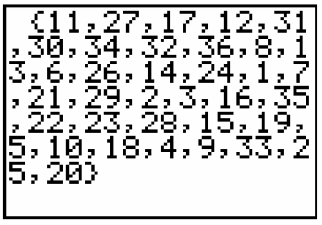


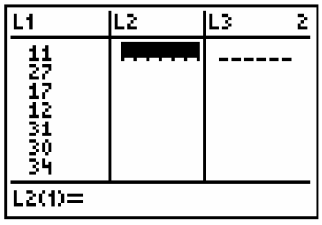


Looking at the values in the first screen, subjects numbered 19, 32, 22, 15, 9, 20, 33, 6, and 2 will receive the first treatment. Subjects 16, 12, 14, 10, 31, 23, 34, 13, and 30 will receive the second treatment. The third treatment will be assigned to subjects 4, 5, and 28. Having generated more sets, we'll complete the treatment 3 group with subjects 18, 29, 24, 11, 35, and 25. Treatment 4 will be assigned to the previously unassigned subjects 1, 3, 7, 8, 17, 21, 26, 27, and 36. While it's fairly easy to do this way, the program can make life a little simpler since there will be no duplicates.

Program Solution. We execute the RANDOM program by numbering the subjects from 1 to 36 and choose all 36.



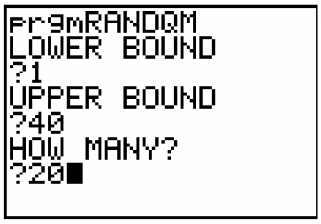


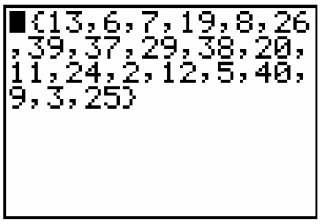


The 36 subjects have been permuted so that they can be assigned randomly to our four groups. Simply use consecutive groups of nine for these groups: {11, 27, 17, 12, 31, 30, 34, 32, 36}, {8, 13, 6, 26, 14, 24, 1, 7, 21}, {29, 2, 3, 16, 35, 22, 23, 28, 15}, {19, 5, 10, 18, 4, 9, 33, 25, 20}.

Example 3.5 More Treatment Assignments. For an experiment with two treatments, we will randomly choose 20 subjects from a group of 40 to receive treatment A. The remaining subjects will receive treatment B.

Solution. We label the subjects from 1 to 40 and then randomly choose 20 for treatment A.



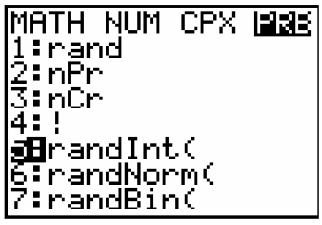


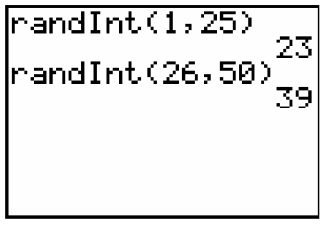
3.3 Sampling Design

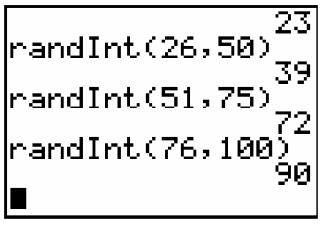
The **RANDOM** program can be used to choose a simple random sample from a designated population. This program chooses the sample all at once without repeated choices. But instead, we may want to use a systematic random sample by drawing one subject at a time from sequential groups. The following exercise demonstrates this process.

Example 3.6 Systematic Sampling. Choose a systematic random sample of four addresses from a list of 100.

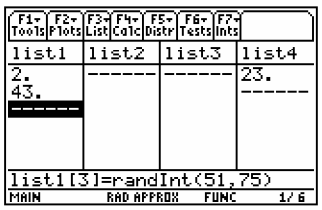
TI-83/84 Solution. Because the list of 100 divides evenly into four groups of 25, we will choose one address from each of the groups 1–25, 26–50, 51–75, and 76–100. And because we are choosing only one number at a time, we can use the `randInt(` command from the `MATH PRB` menu. To choose one integer from *a* to *b*, enter the command `randInt(a,b)`. We use the command four times as shown below to obtain the desired sample.







TI-89 Solution. Place the cursor in the first entry slot in a list. Press [F4] (`Calc`), [4] to expand the `Probability` menu, and [5] to select the `randInt` function. Enter the parameters (low number to select, high number to select), close the parentheses and press [ENTER]. You can repeat these steps changing the low and high numbers to generate the desired four random numbers as above.

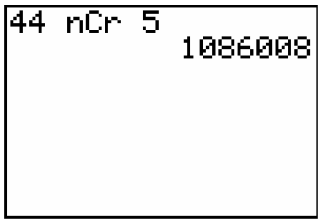


Number of Ways to Choose

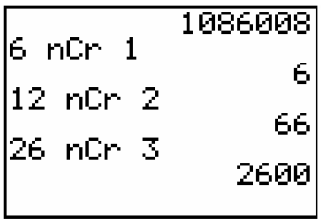
When choosing a simple random sample of size r from a population of size n, we generally choose without repeats and without regard to order. Such a choice is called a *combination*. The number of possible combinations (often called "n choose r") can be computed with the `nCr` option from the `MATH PRB` menu. On a TI-89, this is available on the homescreen (2nd 5 for the `Math` menu) or in the `Stats/List Editor` application from the F4 (`Calc`) `Probability` menu.

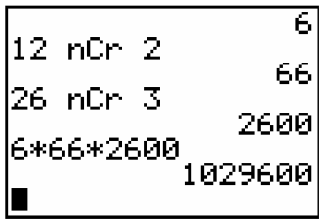
Example 3.7 How Many Samples? (a) How many ways are there to choose five blocks from a group of blocks labeled 1–44? (b) How many ways are there to choose a stratified sample of five blocks so that there is one chosen from blocks 1–6, two chosen from blocks 7–18, and three chosen from blocks 19–44. (c) In each case, choose such a sample.

Solution. (a) There are "44 choose 5" ways to pick five blocks at random from 44, which is computed by `44 nCr 5` on the calculator. Thus, there are 1,086,008 possible samples in this case.

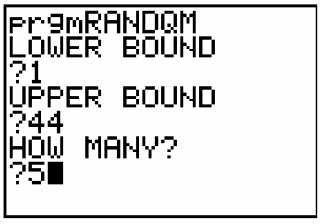


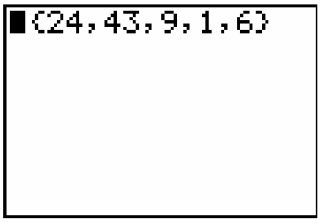
(b) We choose one from the first group of 6 blocks, choose two from the second group of 12 blocks, and choose three from the third group of 26 blocks. The total number of ways to choose in this manner is given by (6 nCr 1)*(12 nCr 2)*(26 nCr 3) = 6 * 66 * 2600 = 1,029,600.

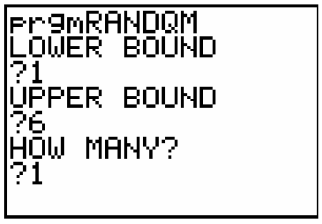


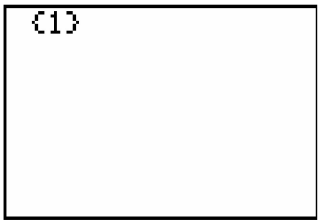


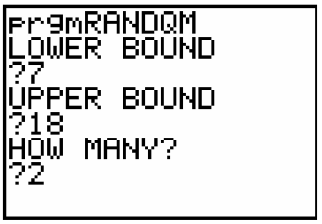
(c) We can choose the samples in each case with the **RANDOM** program. We do not show the last selection.

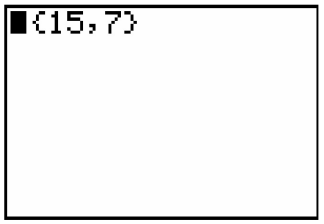










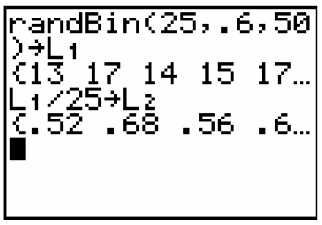


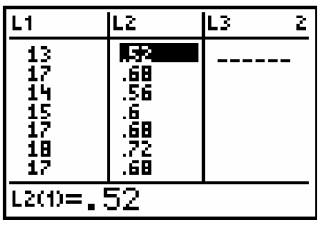
3.4 Toward Statistical Inference

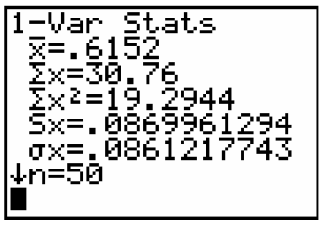
In Section 3.1, we used the command `randBin(1,p,n)`STO▸`L1` to generate a random sample of "Yes/No" responses. In this section, we will demonstrate how to simulate the collection of multiple samples. In particular, we are concerned with the total number of "Yes" responses, the sample proportion for each sample, and the resulting average of all sample proportions. This simulation also can be made with the `randBin(` command from the `MATH PRB` menu.

Example 3.8 Simulating Coin Flips. (a) We have a coin for which the probability of heads is 0.60. We toss the coin 25 times and count the number of heads in this sample. Then we repeat the process for a total of 50 samples of size 25. Simulate the counts of heads for these 50 samples of size 25, compute the sample proportion for each sample, and make a histogram of the sample proportions.

Solution. One simulated sample of counts can be obtained with the command `randBin(25,Ø.6)`. But because we want 50 samples of size 25, we will use the command `randBin(25,Ø.6,5Ø)`STO▸`L1` in order to generate the counts and store them into list `L1`. Then the command `L1/25`STO▸`L2` will compute the sample proportion for each sample and store the results in list `L2`. By computing the sample statistics on list `L2`, we obtain the average of all the sample proportions. This example works the same way on a TI-89; however, you must be in the `Stats/List Editor` Application and have a list name highlighted to issue the `randBin` command. Similarly, highlight list name `list2`, for example, to perform the division.

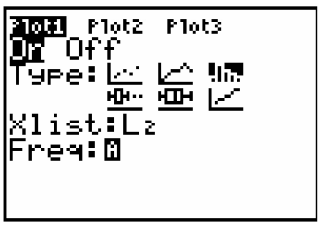


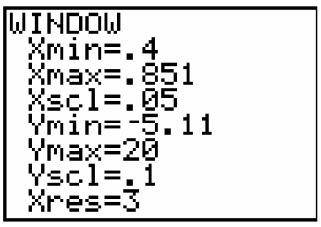


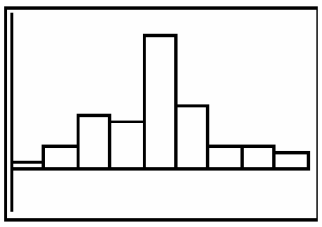


By observing the list of sample proportions in `L2`, we see that we may almost never have a sample proportion that equals the real proportion of 0.60. However, the *average* of all 50 sample proportions was 0.6152, which is very close to 0.60. Lastly, adjust the `WINDOW` and `STAT PLOT` settings to see a histogram of the sample proportions. You

could use ZOOM 9 here to display the histogram, but remember the intervals taken by the calculator are usually not logical. Start there, and then adjust Xmin and Xscl (the bar width) to something that makes sense.

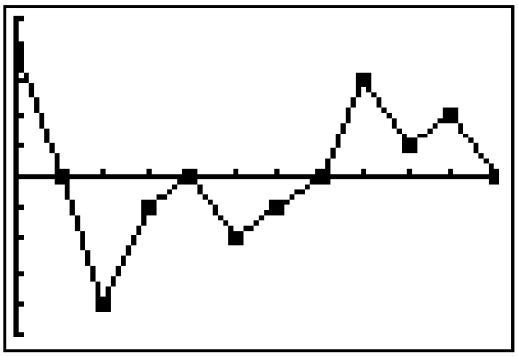






CHAPTER

4



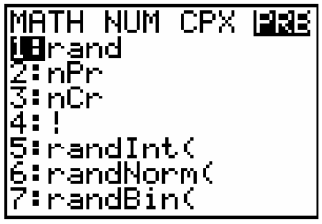
Probability: The Study of Randomness

4.1	Randomness
4.2	Probability Models
4.3	Random Variables
4.4	Means and Variances of Random Variables
4.5	General Probability

Introduction

In this chapter, we show how to use TI calculators to generate some random sequences. We then see how to make a probability histogram for a discrete random variable and how to compute its mean and standard deviation. We conclude with a program for the Law of Total Probability and Bayes' Rule.

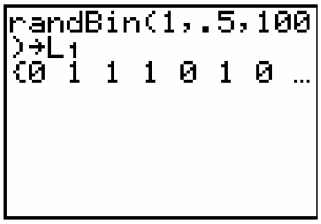
4.1 Randomness

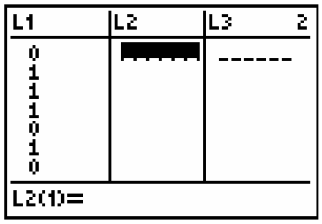


In this section, we work some exercises that use the calculator to generate various random sequences. We shall need the `randBin(` and `randInt(` commands from the `MATH PRB` menu.

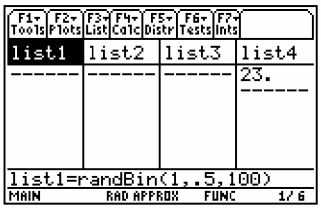
Example 4.1 Simulating Free Throws. Simulate 100 free throws shot independently by a player who has 0.5 probability of making a single shot. Examine the sequence of hits and misses.

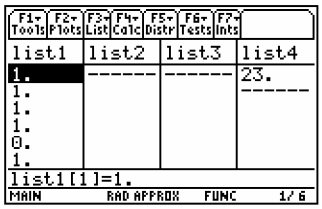
Solution. This is similar to Example 3.1 of the preceding chapter. The command `randBin(1,.5,100)` STO▸ `L1` will generate and store a list of 100 "1s and 0s" to represent the hits and misses. The results can be viewed on the **STAT EDIT** screen.





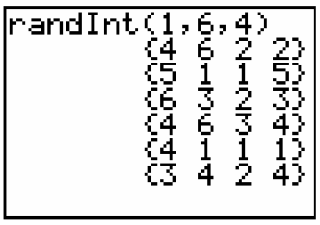
On a TI-89 the solution is similar, but enter the command with the cursor highlighting a list name in the `Stats/List Editor` application. Find the randBin command in the F4 (`Calc`), `Probability` menu.

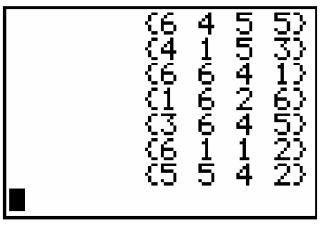


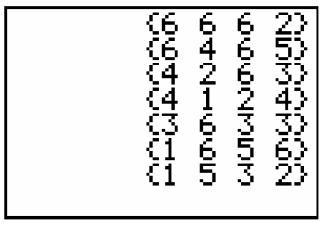


Example 4.2 Simulating Dice. Simulate rolling four fair dice over and over again. What percentage of the time was there at least one "6" in the set of four rolls?

TI-83/84 Solution. The command `randInt(j,k)` generates a random integer from *j* to *k*. The command `randInt(j,k,n)` generates *n* such random integers. Here we enter the command `randInt(1,6,4)` to simulate four rolls of dice numbered 1 to 6. After entering the command once, keep pressing ENTER to reexecute the command.

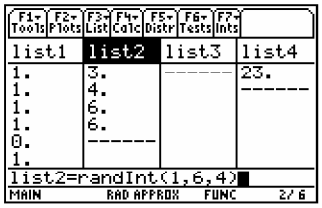






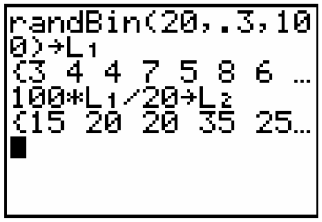
In the 20 sets shown above, there are 13 sets with at least one "6," which gives an estimate of 65%.

TI-89 Solution. This is not as elegant as with the other models. The TI-89 will only generate random numbers of this type when the command is entered while a list name is highlighted. Keep reentering the command to repeat the simulation of rolling four dice.

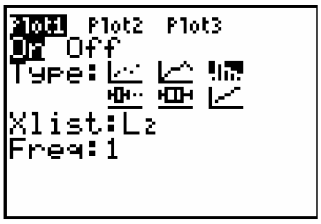


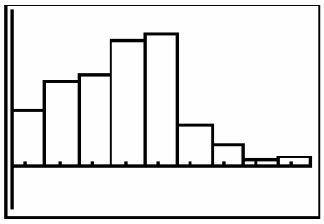
Example 4.3 Simulating Binomial Counts. Simulate 100 binomial observations each with $n = 20$ and $p = 0.3$. Convert the counts into percents and make a histogram of these percents.

Solution.
The command `randBin(20,.3,100)` STO▸ `L1` will generate the observations and put them in list `L1`, and the command `100*L1/20` STO▸ `L2` will put the percents into list `L2`. On a TI-89, enter the commands with a list name highlighted in the `Stats/List Editor` application.



Next, adjust the `STAT PLOT` settings to make a histogram of the percents in list `L2`. If necessary, use `WINDOW` to adjust the plot for logical intervals.



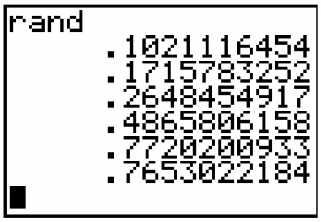


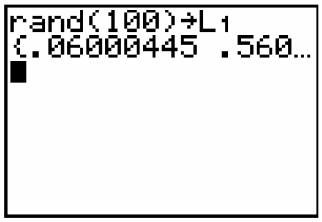
4.2 Probability Models

In this section, we demonstrate some of the basic concepts of probability models.

Example 4.4 Generating Random Numbers (a) Generate random numbers between 0 and 1. (b) Make a histogram of 100 such randomly generated numbers.

Solution. (a) Simply enter the command `rand` from the `MATH PRB` menu. On a TI-89, this can be done either on the home screen, or in the Stats/List Editor application. To store a list on the TI-89, use the command with the list name highlighted as has been shown before. After the command has been entered once, keep pressing ENTER to continue generating more random values between 0 and 1. (b) Enter the command `rand(100)` STO▸ `L1` to generate and store 100 random numbers between 0 and 1.



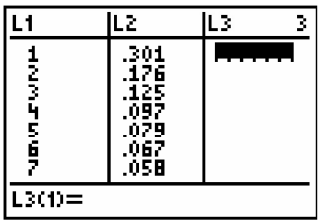


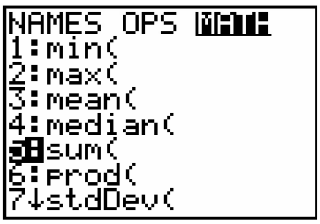
Example 4.5 Benford's Law. The first digit v of numbers in legitimate records often follows the distribution given in the table below, known as Benford's Law. (a) Verify that the table defines a legitimate probability distribution. (b) Compute the probability that the first digit is 6 or greater.

First digit v	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Solution. (a) To sum these probabilities, we can enter the values into a list and use the command `sum(` from the `LIST MATH` menu.

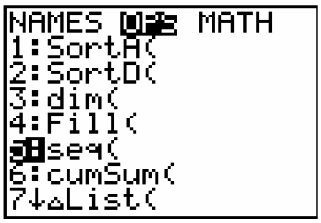
First, enter the values of the digits into list `L1` (optional at this point, but we'll use it later) and enter the probabilities into list `L2`. To verify that the table gives a legitimate probability distribution, enter the command `sum(L2)`, which sums the values in list `L2`. If you are using a TI-89, move the cursor to an empty list entry, and find the command in the F3 `List` menu.

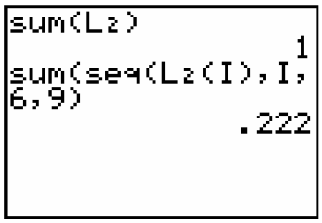






(b) For this part of the question, we only want to sum some of the entries. After summing all the probabilities, press 2nd ENTER to recall the previous command, then edit it to `sum(seq(L2(I),I,6,9))` to sum the sixth through ninth probabilities. We find the `seq(` command shell on the `LIST OPS` menu. We see that the probability of the first digit being 6 or greater is 0.222. You will need to press the ALPHA key in order to type the I's in the command.





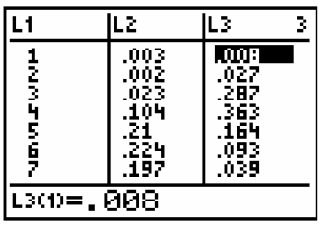
4.3 Random Variables

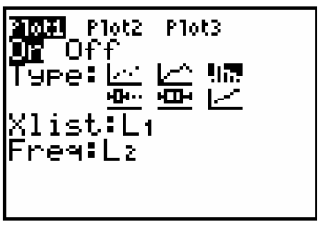
In this section, we work on exercises that compute various probabilities involving random variables. We begin with an exercise on constructing a probability histogram.

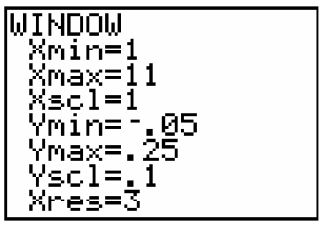
Example 4.6 How Many Rooms? The table below gives the distributions of rooms for owner-occupied units and for renter-occupied units in San Jose, California. Make probability histograms of these two distributions. If X represents the number of rooms in a randomly chosen owner-occupied unit, compute $P(X > 5)$.

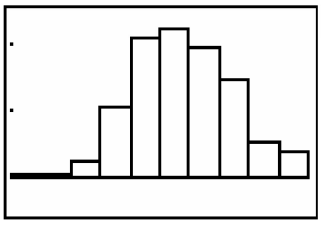
Rooms	1	2	3	4	5	6	7	8	9	10
Owned	0.003	0.002	0.023	0.104	0.210	0.224	0.197	0.149	0.053	0.035
Rented	0.008	0.027	0.287	0.363	0.164	0.093	0.039	0.013	0.003	0.003

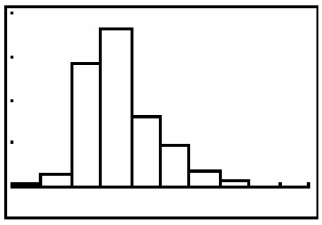
Solution. We enter the values of the rooms into list L1, the owner probabilities into list L2, and the renter probabilities into list L3. Then we make separate histograms for an Xlist of L1 with frequencies of either L2 or L3. Since the number of rooms is discrete, we'll set the window so each bar has width 1. The left histogram is for the owned units, the right for rented units. Note: We had to change Ymax for the rented units, as their largest frequency is higher than in the owned distribution.









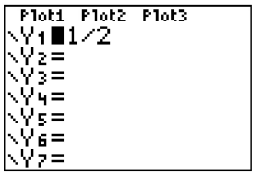


After displaying these histograms, you will need to set Freq: on the plot definition screen back to a 1 by pressing ALPHA 1.

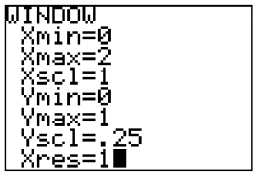
The value $P(X > 5)$ is equivalent to $P(6 \le X \le 10)$. This value can be computed with the command sum(seq(L2(I),I,6,10)) as described above, which gives 0.658.

Example 4.7 Uniform Distributions. Let $Y \sim U[0, 2]$. (a) Graph the density curve. (b) Find $P(0.5 < Y < 1.3)$.

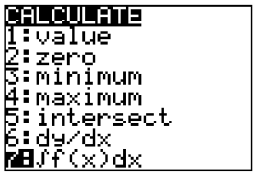
Solution. For $Y \sim U[0, 2]$, the height of the density curve is $1/(2 - 0) = 0.5$. We simply enter this function into the Y= screen and use item 7 from the CALC menu to compute the area between the values 0.5 and 1.3. The steps are similar on a TI-89 (using the top row buttons preceded by ◆). Use the F5 (Math) button to find the numerical integration command.



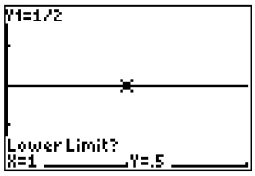
Enter 1/2 into Y= screen.



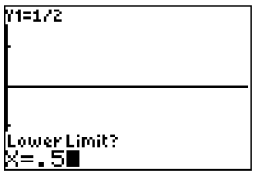
Set WINDOW with X from 0 to 2.



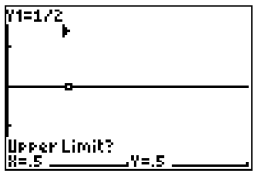
Press CALC (2nd TRACE), then 7.



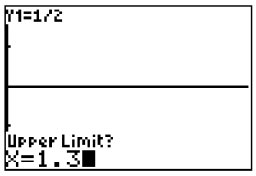
When screen appears, type .5



Press ENTER after typing .5.



When new screen appears, type 1.3.



Press ENTER after typing 1.3.

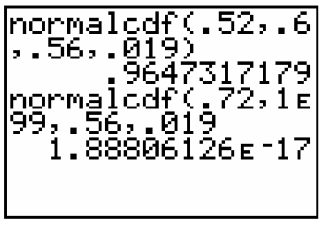


$P(0.5 \le Y \le 1.3)$ is shown as 0.4.

We conclude this section by working an exercise using the normal density curve that reviews the normal distribution calculations from Section 1.3 of this manual.

Example 4.8 Voting in Oregon. After an election in Oregon, voter records showed that 56% of registered voters actually voted. A survey of 663 registered voters is conducted and the sample proportion $\hat{p}$ of those who claim to have voted is obtained. For all random samples of size 663, these values of the sample proportions $\hat{p}$ will follow an approximate normal distribution with mean $\mu = 0.56$ and standard deviation $\sigma = 0.019$. Use this distribution to compute $P(0.52 < \hat{p} < 0.60)$ and $P(\hat{p} \ge 0.72)$.

Solution. We use the normalcdf(command from the DISTR menu. For $\hat{p} \sim N(.56,.019)$, we can find $P(0.52 < \hat{p} < 0.60)$ = using the command normalcdf(.52,Ø.6,.56,.Ø19). To find $P(\hat{p} \ge 0.72)$, use normalcdf(.72,1E99,.56,.Ø19), although practically any large positive value (such as 999999) would do. We find there is a 96.5% chance that between 52% and 60% would say they had voted. There is practically no chance that more than 72% would say they had voted.



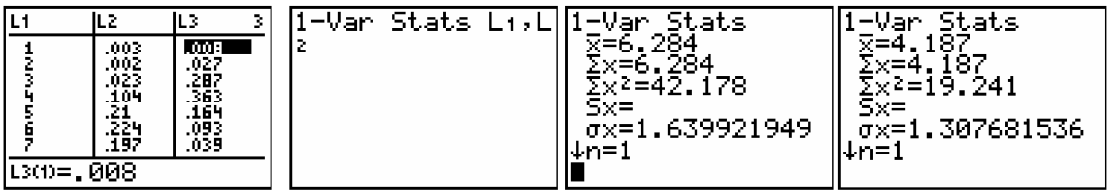
4.4 Means and Variances of Random Variables

We now show how to compute the mean and standard deviation of a discrete random variable for which the range of measurements and corresponding probabilities are given.

Example 4.9 More About Rooms. The table below gives the distributions of the number of rooms for owner-occupied units and renter-occupied units in San Jose, California, which we saw earlier. We want to calculate the mean and the standard deviation of the number of rooms for each type.

Rooms	1	2	3	4	5	6	7	8	9	10
Owned	0.003	0.002	0.023	0.104	0.210	0.224	0.197	0.149	0.053	0.035
Rented	0.008	0.027	0.287	0.363	0.164	0.093	0.039	0.013	0.003	0.003

Solution. If they are not still there, enter the measurements (rooms) into list L1 and the probabilities into lists L2 and L3. For the owner-occupied units, enter the command 1-Var Stats L1,L2. We see that the average number of rooms for owner-occupied units is $\mu = 6.284$ with a standard deviation of $\sigma = 1.64$. For the renter-occupied units, enter the command 1-Var Stats L1,L3 to obtain $\mu = 4.187$ and $\sigma = 1.3077$. Remember, since this is a probability distribution, we want parameter values. Notice the calculator is smart enough to skip the sample standard deviations!



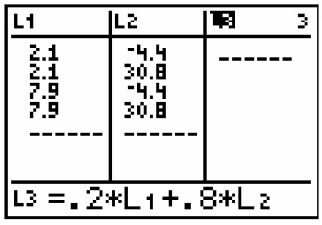
Mean and Standard Deviation of an Independent Sum

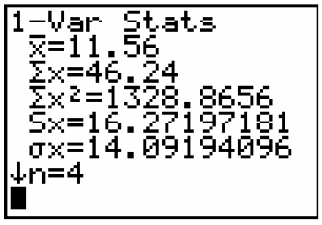
Here we show how to verify the mean and standard deviation of a random variable of the form $Z = aX + bY$, where X and Y are independent.

Example 4.10. Suppose that $Z = 0.2X + 0.8Y$, where $\mu_X = 5$, $\sigma_X = 2.9$, $\mu_Y = 13.2$, and $\sigma_Y = 17.6$. Assuming that X and Y are independent, find the mean and standard deviation of Z.

Solution. We can think of X as taking two values $5 - 2.9 = 2.1$ and $5 + 2.9 = 7.9$. Likewise, we can consider Y to assume only the values $13.2 - 17.6 = -4.4$ and $13.2 + 17.6 = 30.8$. We enter these values of X and Y into lists L1 and L2; however, we list each X value twice in consecutive fashion (2.1, 2.1, 7.9, 7.9), and we list the Y values twice in

alternating fashion (–4.4, 30.8, –4.4, 30.8). Next, we enter the command `Ø.2*L1 + Ø.8*L2` while name `L3` is highlighted to send the possible values of Z to list `L3`. Finally, enter `1-Var Stats L3` to compute the mean and standard deviation of Z.





We thereby can verify that $\mu_Z = 0.2\mu_X + 0.8\mu_Y$ and, because X and Y are independent, that $\sigma_Z = \sqrt{0.2^2\sigma^2{}_X + 0.8^2\sigma^2{}_Y}$. However, if $Z = aX + bY$ and X and Y are *not* independent, then $\sigma_Z = \sqrt{a^2\sigma^2{}_X + b^2\sigma^2{}_Y + 2\rho a\sigma_X b\sigma_y}$.

4.5 General Probability

We conclude this chapter with a short program for the Law of Total Probability and Bayes's rule. The **BAYES** program given below computes and displays the total probability $P(C)$ according to the formula

$$P(C) = P(A_1)P(C \mid A_1) + P(A_2)P(C \mid A_2) + \ldots + P(A_n)P(C \mid A_n)$$

Before executing the **BAYES** program, enter given probabilities $P(A_1), \ldots, P(A_n)$ into list `L1` and the given conditionals $P(C \mid A_1), \ldots, P(C \mid A_n)$ into list `L2`.

The program also stores the probabilities of the intersections, $P(C \cap A_1), \ldots, P(C \cap A_n)$ in list `L3`, and stores the reverse conditionals $P(A_1 \mid C), \ldots, P(A_n \mid C)$ in list `L4`. Finally, the conditional probabilities $P(A_1 \mid C'), \ldots, P(A_n \mid C')$ are stored in list `L5`, and the conditional probabilities $P(C \mid A_1'), \ldots, P(C \mid A_n')$ are stored in list `L6`. To help the user keep track of which list contains which probabilities, the program displays a description upon completion.

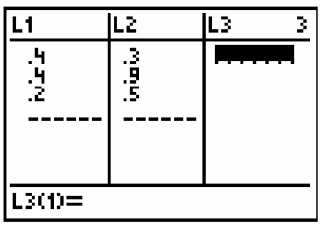
The program, with appropriate list name changes, will also run on a TI-89.

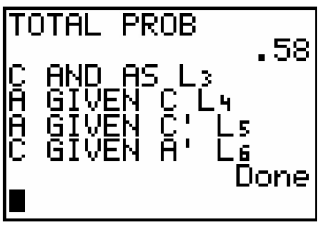
The BAYES Program

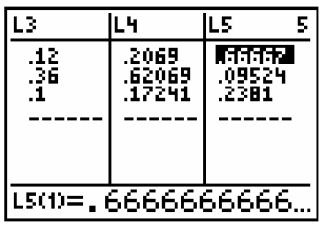
PROGRAM:BAYES	:Disp "TOTAL PROB"
:L1*L3 STO▶ L3	:Disp round(T,4)
:sum(L3) STO▶ T	:Disp "C AND As : L3"
:L3/T STO▶ L4	:Disp "As GIVEN C : L4"
:L1*(1-L2)/(1-T) STO▶ L5	:Disp "As GIVEN C' : L5"
:T*(1-L4)/(1-L1) STO▶ L6	:Disp "C GIVEN A's : L6"

Example 4.11 Will He Win? The voters in a large city are 40% white, 40% black, and 20% Hispanic. A mayoral candidate expects to receive 30% of the white vote, 90% of the black vote, and 50% of the Hispanic vote. Apply the **BAYES** program to compute the percent of the overall vote that the candidate expects, and to analyze the other computed conditional probabilities.

Solution. Here we let A_1 = white voters, A_2 = black voters, and A_3 = Hispanic voters. We enter the probabilities of these events into list L1. We let C be the event that a person votes for the candidate. Then $P(C \mid A_1) = 0.30$, $P(C \mid A_2) = 0.90$, and $P(C \mid A_3) = 0.50$, and we enter these conditional probabilities into list L2. Then we execute the **BAYES** program.







We first see that the candidate can expect to receive 58% of the overall vote. This result is obtained by $P(C) = .4*.3 + .4*.9 + .2*.5 = 0.58$. The other computed probabilities are also stored in the designated lists.

List L3 contains the probabilities of the intersections. The probability that a voter is white and will vote for the candidate is $P(A_1 \cap C) = 0.12$. The probability that a voter is black and will vote for the candidate is $P(A_2 \cap C) = 0.36$. The probability that a voter is Hispanic and will vote for the candidate is $P(A_3 \cap C) = 0.1$. These values are obtained with the formula $P(A_i \cap C) = P(A_i)P(C \mid A_i)$, and are found by multiplying the terms in lists L1 and L2.

List L4 contains the reverse conditional probabilities of being white, black, Hispanic given that one *will* vote for the candidate. This list is the direct application of Bayes's rule. These conditional probabilities are $P(A_1 \mid C) = 0.2069$, $P(A_2 \mid C) = 0.62069$, and

$P(A_3 \mid C)$ = 0.17241, respectively. This list is obtained by dividing the respective intersection probabilities in list L3 by $P(C) = 0.58$.

List L5 contains the reverse conditional probabilities of being white, black, Hispanic given that one will *not* vote for the candidate. These values are $P(A_1 \mid C') = 0.66667$, $P(A_2 \mid C') = 0.09524$, and $P(A_3 \mid C') = 0.2381$, respectively.

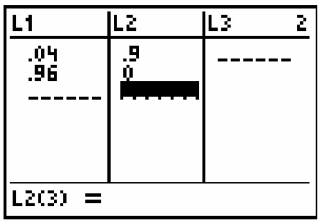
List L6 (not shown) contains the conditional probabilities of voting for the candidate given that a voter is *not* white, *not* black, and *not* Hispanic. These values are $P(C \mid A_1') = 0.76667$, $P(C \mid A_2') = 0.36667$, and $P(C \mid A_3') = 0.6$, respectively.

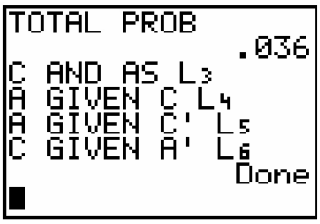
We note that given any conditional probability $P(C \mid D)$, then the complement conditional probability is given by $P(C' \mid D) = 1 - P(C \mid D)$. Thus, lists for complement conditional probabilities do not need to be generated. For example, the respective conditional probabilities of *not* voting for the candidate given that one is white, black, and Hispanic are respectively 0.70, 0.10, and 0.50. These values are the complement probabilities of the originally given conditionals.

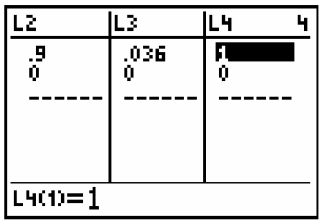
Example 4.12 Cystic Fibrosis. The probability that a randomly chosen person of European ancestry carries an abnormal CF gene is 1/25. If one is a carrier of this gene, then a test for it will be positive 90% of the time. If a person is not a carrier, then the test will never be positive. Jason tests positive. What is the probability that he is a carrier?

Solution. We let C be the event that a person of European ancestry tests positive, let A_1 be carriers of the gene, and let A_2 be noncarriers. Because $P(A_1)$ = 1/25, then $P(A_2) =$ 24/25. Also, $P(C \mid A_1) = 0.90$, $P(C \mid A_2) = 0$. We are trying to compute $P(A_1 \mid C)$, which is the probability of being a carrier given that one has tested positive.

We enter the probabilities $P(A_1)$ and $P(A_2)$ into list L1, enter the conditional probabilities $P(C \mid A_1)$ and $P(C \mid A_2)$ into list L2, and execute the **BAYES** program.





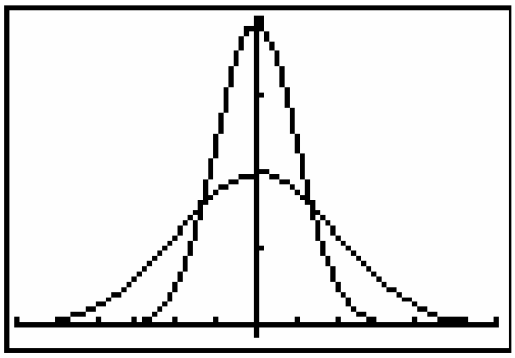


List L4 contains the conditional probabilities $P(A_i \mid C)$. From the first entry, we see that

$P(A_1 \mid C) = 1$. Because Jason has tested positive, there is a 100% chance that he is a carrier. (He must be a carrier because it is impossible for noncarriers to test positive.) This value is also given by $P(A_1 \mid C) = \dfrac{P(A_1 \cap C)}{P(C)} = \dfrac{1/25 * 0.90}{1/25 * .90 + 24/25 * 0} = 1$.

CHAPTER

5



Sampling Distributions

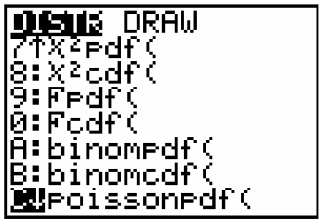
5.1 Sampling Distributions for Counts and Proportions
5.2 Poisson Random Variables
5.3 The Sampling Distribution of a Sample Mean
5.4 Common Errors

Introduction

In this chapter, we show how to compute probabilities involving binomial distributions, Poisson distributions, and the sample mean $\overline{x}$.

5.1 Sampling Distributions for Counts and Proportions

We begin by demonstrating how to compute various probabilities for a given binomial distribution. To do so, we will need the `binompdf(` and `binomcdf(` commands from the `DISTR` menu. Actual location of these depends on the calculator model.

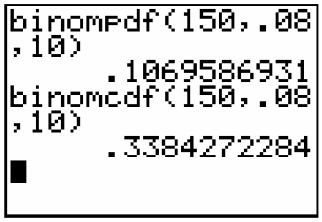


Binomial Probabilities

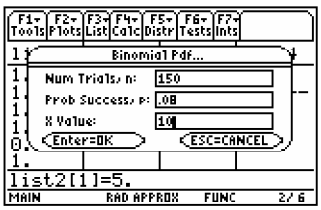
For a binomial distribution $X \sim B(n, p)$, we compute the probability of exactly k successes, $P(X = k)$, by entering the command `binompdf(n,p,k)`. The probability $P(X \le k) = P(0 \le X \le k)$ of at most k successes is computed with the command `binomcdf(n,p,k)`. The probability of there being at least k successes is given by $P(X \ge k) = 1 - P(X \le k-1)$, and is computed with the command `1-binomcdf(n,p,k-1)`. The following three examples demonstrate these calculations. TI-89 calculators explicitly ask for the low and high ends of interest to be included in the calculation, so there is no need for any subtraction.

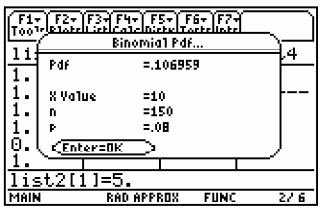
Example 5.1 Auditing Sales. An audit examines a simple random sample of 150 out of 10,000 available sales records. Suppose that in fact, 800 of the 10,000 sales are incorrectly classified. What is the probability we find *exactly* 10 misclassified records? What is the probability we find *at most* 10 misclassified records? Since 800/10000 = 0.08, we let X = the number of misclassified records, and $X \sim$ B(150, 0.08). Calculate $P(X = 10)$ and $P(X \le 10)$.

TI-83/84 Solution. Simply enter the commands `binompdf(150,.08,10)` and `binomcdf(150,.08,10)` to obtain $P(X = 10) \approx 0.1070$ and $P(X \le 10) \approx 0.3384$. As always, ask your instructor how many decimal places to report.

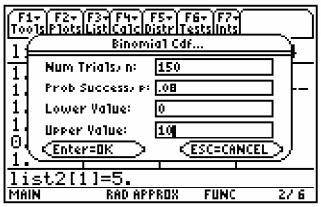


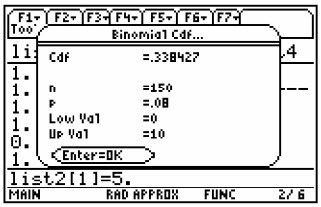
TI-89 Solution. In the Stats/List Editor application, press F5 for the Distr menu. Arrow down to find Binomial Pdf. Press ENTER to select it. You will have a dialog box like the one below that explicitly asks for n, p, and k. Press ENTER to perform the calculation.





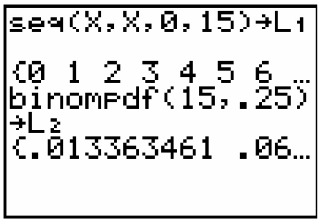
`Binomial Cdf` is similar, however. In this case, we specify the low and high ends of interest.

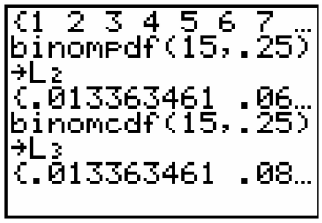


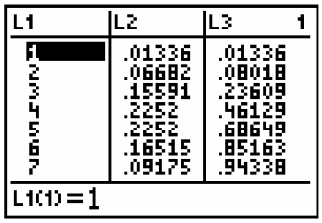


Example 5.2 Guessing on a Quiz. Suppose a multiple choice quiz has 15 questions, each with 4 possible answers. If a student is purely guessing, what is the probability of at most one correct answer? Let $X \sim B(15, 0.25)$. Make a probability table and probability histogram of the distribution. Also make a table of the cumulative distribution and use it to find $P(X \le 1)$.

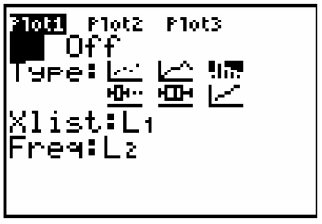
Solution. Because there are $n = 15$ attempts, the possible number of successes range from 0 to 15. So we first enter the integer values 0, 1, . . . , 15 into list `L1`. We can do so directly or we can use the command `seq(X,X,0,15)` from the `LIST` menu with list name `L1` highlighted by the cursor. Next, we use the commands `binompdf(15,.25)` `STO▶` `L2` and `binomcdf(15,.25)` `STO▶` `L3` to enter the probability distribution values $P(X = k)$ into list `L2` and the cumulative distribution values $P(X \le k)$ into list `L3`. If you are using a TI-89, enter all these commands with the receiving list name highlighted by the cursor. From the values in list `L3`, we can see that $P(X \le 1) \approx 0.0802$.

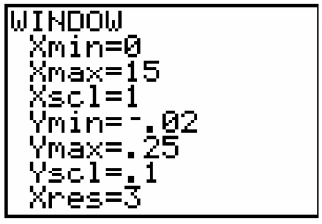


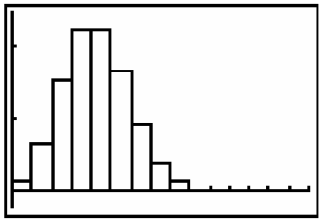




To see a probability histogram of the distribution, adjust the `STAT PLOT` settings for a histogram of `L1` with frequencies `L2`. We note that the probabilities beyond $X = 9$ will not be observable, because those probabilities are so small. Once again, for this histogram, since we want bars of width 1, the `ZOOM` `9` picture will need adjusting for an `Xscl` of 1. I have shown my final `WINDOW` settings below.

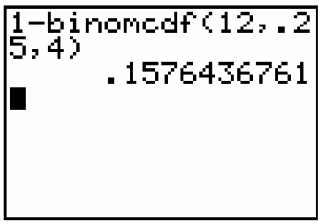






Example 5.3 Free Throws. Suppose a basketball player makes 75% of his free throws. In one particular game, he missed 5 of 12 attempts. Is it unusual to perform this poorly? We let X = number of missed free throws, so X ~ B(12,.25), and compute $P(X \geq 5)$.

Solution. We use the probability of the complement to obtain $P(X \geq 5) = 1 - P(X \leq 4)$, which is computed using `1-binomcdf(12,.25,4)` ≈ 0.1576. If you are using a TI-89, simply enter a low end of 5 and a high end of 12.

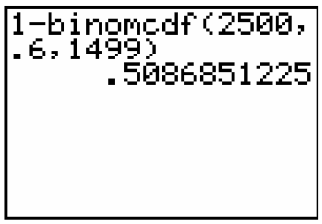


Probabilities for $\hat{p}$

The next two examples show how to make probability calculations for a sample proportion $\hat{p}$ by converting to a binomial probability.

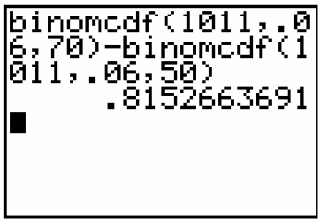
Example 5.4 Clothes Shopping. Suppose that 60% of all adults agree that they like shopping for clothes, but often find it frustrating and time-consuming. In a nationwide sample of 2500 adults, let $\hat{p}$ be the sample proportion of adults who agree with this response. Compute $P(\hat{p} \geq 0.58)$, the chance that more than 58% in your sample will agree that clothes shopping can be frustrating and time-consuming.

Solution. Because 58% of 2500 is 1450, we must compute $P(X \geq 1450)$, where $X \sim B(2500, 0.60)$. Instead, we may compute $1 - P(X \leq 1499)$ using `1-binomcdf(2500,.6,1499)` ≈ 0.5087.



Example 5.5 Betting on Football. A Gallup poll of size n = 1011 found 6% of the respondents said they bet on college football. Assuming a true proportion of p = 0.06, what is the probability that a sample proportion $\hat{p}$ lies between 0.05 and 0.07?

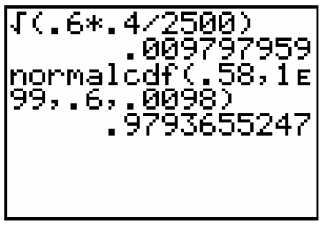
Solution. For n = 1011 and p = 0.06, then $P(0.05 < \hat{p} < 0.07) = P(.05*n < X < .07*n) = P(50.55 < X < 70.77)$. Because there can't be fractions of "successes" this becomes $P(51 \leq X \leq 70) = P(X \leq 70) - P(X \leq 50)$, where X ~ B(1011, 0.06). We find this value by entering `binomcdf(1011,.06,70)-binomcdf(1011,.06,50)`, and obtain a probability of about 0.8153.



Normal Approximations

We conclude this section by showing how to approximate a sample proportion probability and a binomial probability with a normal distribution.

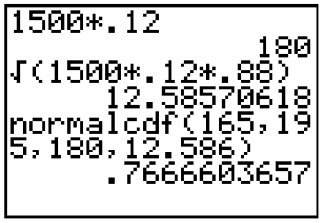
Example 5.6 More on Clothes Shopping. With n = 2500 and p = 0.60 as in Example 5.4 above, use the approximate distribution of $\hat{p}$ to estimate $P(\hat{p} > 0.58)$.



Solution. The distribution of $\hat{p}$ is approximately normal with $\mu = p = 0.60$ and
$\sigma = \sqrt{p(1-p)/n} = \sqrt{.6*.4/2500} = 0.0098$.
Thus, $P(\hat{p} \geq 0.58) \approx \mathrm{P}(Y \geq 0.58)$, where Y ~ N(0.60, 0.0098).
The command `normalcdf(.58,1E99,.6,.0098)` gives a probability of about 0.9794.

Example 5.7 Checking for Survey Errors. One way of checking the effect of undercoverage, nonresponse, and other sources of error in a sample survey is to compare the sample with known facts about the population. About 12% of American adults are black. The number X of blacks in a random sample of 1500 should therefore be binomial with n = 1500 and p = 0.12. (a) What are the mean and standard deviation of X?
(b) Use the Normal approximation to find the chance there are between 165 and 195 blacks in our survey.

Solution. (a) The mean is $\mu = np = 1500*0.12 = 180$, and the standard deviation is
$\sigma = \sqrt{np(1-p)} = \sqrt{1500*.12*.88} = 12.586.$



(b) We now let Y ~ N(180,12.586). Then $P(165 < X < 195)$ is found using `normalcdf(165,195,180,`12.586`)`. We see that the desired probability $\approx$ 0.7667.

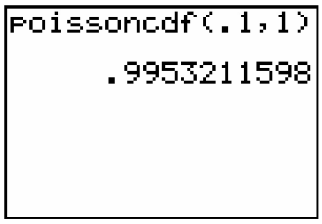
5.2 Poisson Random Variables

Poisson random variables generally come from one of two situations. We have a binomial random variable with large n and small p so that the expected number of "successes" is small, or we have a process, such as telephone calls to a switchboard, where we only observe the successes. Poisson random variables have only one parameter, λ, the average "rate" of the process. In the binominal setting, $\lambda = np$.

Example 5.8 Mumps Outbreaks. Mumps is an acute viral infection that is generally mild, and in about 20% of infected individuals, even asymptomatic (meaning they do not know they have the disease). However, severe complications can arise. Mandatory vaccinations have largely eradicated the disease in the United States. For the state of

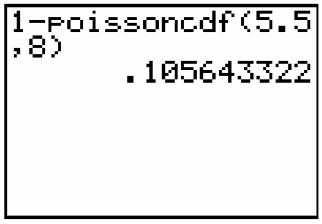
Iowa, the average monthly number of reported cases is about 0.1 per month. Assuming that cases are independent, what is the probability that in a given month, there will be no more than one case of mumps in Iowa?

Solution. The rate of the process is $\lambda = 0.1$ per month. We let X be the number of cases in a month and want to know $P(X \leq 1)$. Just as in the binomial case, this is a cumulative probability (the probability of either 0 or 1 case of mumps), so we use Poissoncdf(from the DISTR menu. Parameters are λ and k. The probability of at most one case of mumps per month in Iowa is about 0.9953.



Example 5.9 ATM Use. Suppose the number of people who use a certain ATM machine can be modeled as a Poisson process. The average number of users is 5.5 per hour. What is the probability that in a given hour, more than 8 will use the machine?

Solution. The rate of the process is $\lambda = 5.5$ per hour. If X is the number of users of the machine in an hour, we want to find $P(X > 8)$. Just as with binomial probabilities, we find this probability using the complements rule as $1 - P(X \leq 8)$. The probability of more than 8 ATM users in an hour for this machine is about 10.56%.



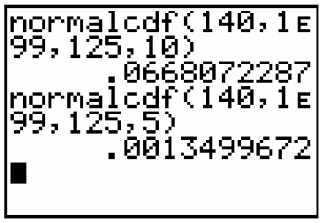
5.3 The Sampling Distribution of a Sample Mean

We now show how to compute various probabilities involving the sample mean $\bar{x}$. To do so, we make use of the fact that for random samples of size n from a $N(\mu, \sigma)$ distribution, the sample mean $\bar{x}$ follows a $N(\mu, \sigma/\sqrt{n})$ distribution. According to the Central Limit Theorem, when samples are "large," $\bar{x}$ also follows a $N(\mu, \sigma/\sqrt{n})$ distribution.

Example 5.10 Measuring Blood Glucose. Sheila's glucose level one hour after ingesting a sugary drink varies according to the Normal distribution with μ = 125 mg/dl and σ = 10 mg/dl.

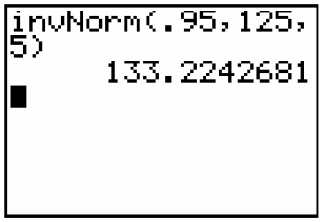
(a) If a single glucose measurement is made, what is the probability that Sheila measures above 140?
(b) What is the probability that the sample mean from four separate measurements is above 140?

Solution. (a) We compute $P(X > 140)$ for $X \sim N(125,10)$ with the command `normalcdf(140,1E99,125,10)`. Then, $P(X > 140) \approx 0.0668$.
(b) For an SRS of size $n = 4$, $\bar{x}$ has a mean of $\mu = 125$ and a standard deviation of $\sigma/\sqrt{n} = 10/\sqrt{4} = 5$. So now we compute $P(\bar{x} > 140)$ for $\bar{x} \sim N(125, 5)$ and obtain a value of about 0.00135.



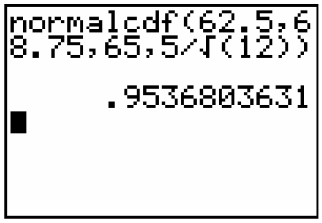
Example 5.11 More Blood Glucose. Sheila's glucose level one hour after ingesting a sugary drink varies according to the normal distribution with $\mu = 125$ mg/dl and $\sigma = 10$ mg/dl. What is the level L such that there is only 0.05 probability that the mean glucose level of four test results falls above L for Sheila's glucose level distribution?

Solution. As in the previous example, $\bar{x} \sim N(125, 5)$. So we must find the inverse normal value L for which $P(\bar{x} > L) = 0.05$ or, equivalently, $P(\bar{x} \le L) = 0.95$. We compute this value with the `invNorm(` command from the `DISTR` menu by entering `invNorm(.95,125,5)`. We see that only about 5% of the time should $\bar{x}$ be larger than $L = 133.224$.



Example 5.12 Egg Weights. The weight of eggs produced by a certain breed of hen is normally distributed with a mean of 65 g and a standard deviation of 5 g. For random cartons of 12 eggs, what is the probability that the weight of a carton falls between 750 g and 825 g?

Solution. If the total weight of 12 eggs falls between 750 g and 825 g, then the sample mean $\bar{x}$ falls between 750/12 = 62.5 g and 825/12 = 68.75 g. So, we compute $P(62.5 < \bar{x} < 68.75)$ for $\bar{x} \sim N(65, 5/\sqrt{12})$ using the command `normalcdf(62.5,68.75,65,5/√(12))`. Note: If you are using a TI-89, you will first have to find $\sigma(\bar{x}) = 5/\sqrt{12} = 1.443$.



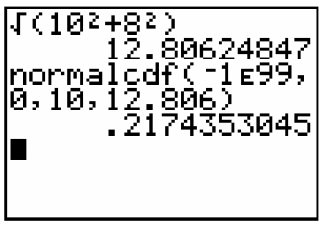
Sum of Independent Normal Measurements

Let $X \sim N(\mu_X, \sigma_X)$ and let $Y \sim N(\mu_Y, \sigma_Y)$. Assuming that X and Y are independent measurements, then $X \pm Y$ follows a $N\left(\mu_X \pm \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right)$ distribution.

Example 5.13 A Golf Competition. Tom and George are playing in the club golf tournament. Their scores vary as they play the course repeatedly. Tom's score X has the

$N(110, 10)$ distribution and George's score Y has a $N(100, 8)$ distribution. If the scores are independent, then what is the probability that Tom's score is less than George's?

Solution. The value $P(X < Y)$ is equivalent to $P(X - Y < 0)$. So we need to use the distribution of the difference $X - Y$ which is $N(110 - 100, \sqrt{10^2 + 8^2}) = N(10, 12.806)$. Using the command `normalcdf(-1E99,0,10,12.806)`, we find that $P(X - Y < 0) \approx 0.2174$. Although George has a lower average, Tom should beat him about 21% of the time.

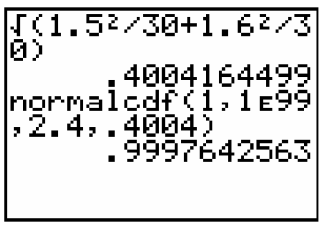


Sum and Difference of Sample Means

Let $\bar{x}$ be the sample mean from an SRS of size n from a $N(\mu_X, \sigma_X)$ distribution, and let $\bar{y}$ be the sample mean from an independent SRS of size m from a $N(\mu_Y, \sigma_Y)$ distribution. Then, the sum/difference $\bar{x} \pm \bar{y}$ follows a $N\left(\mu_X \pm \mu_Y, \sqrt{\sigma_X^2 / n + \sigma_Y^2 / m}\right)$ distribution.

Example 5.14 Credible Sources. In a randomized comparative experiment, students read ads that cited either the *Wall Street Journal* or the *National Enquirer*. They were asked to rate the trustworthiness of the source on a 7 point scale. Let $\bar{y}$ be the sample mean from the group who read the ad citing the *Journal* and $\bar{x}$ be the sample mean from a group of size 30 who read the ad citing the *Enquirer*. Suppose the population of all student scores for the *Journal* have a $N(4.8, 1.5)$ population, and the population of all student scores for the *Enquirer* are $N(2.4, 1.6)$. What is the distribution of $\bar{y} - \bar{x}$? Find $P(\bar{y} - \bar{x}) \geq 1$.

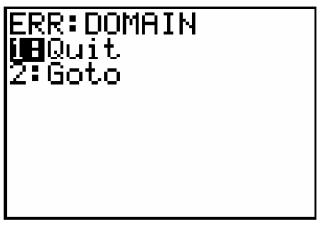
Solution. First, $\bar{y} - \bar{x} \sim N(4.8 - 2.4, \sqrt{\frac{1.5^2}{30} + \frac{1.6^2}{30}}) = N(2.4, 0.4004)$. We find $P(\bar{y} - \bar{x}) \geq 1$ using `normalcdf(1,1E99,2.4,0.4004)` and obtain a value of 0.9998. It's virtually certain the *Journal* will have a higher mean score at least one point higher than the *Enquirer*.



5.4 Common Errors

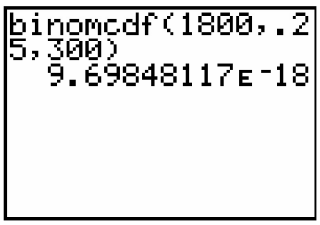
Err:Domain

This error is normally caused in these types of problems by specifying a probability as a number greater than 1 (in percent possibly instead of a decimal) or a value for *n* or *x* that is not an integer. Reenter the command giving *p* in decimal form. Pressing [ESC] will return you to the input screen to correct the error. This will also occur in older TI-83 calculators if *n* is too large in a binomial calculation; in that case, you need to use the normal approximation.



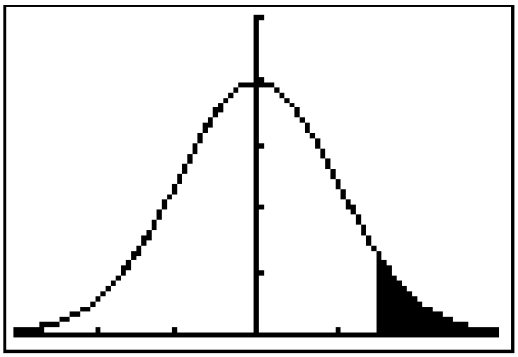
Probabilities larger than 1

This is not possible. As we've said before, if it appears more than 1 on the first glance, check the right hand side. This value is 9.7 x 10^{-18} or 17 0s followed by the leading 9.



CHAPTER

6



Introduction to Inference

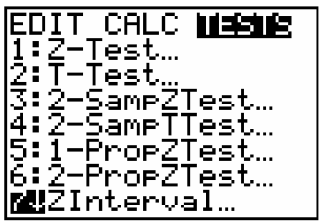
6.1	Confidence Intervals with σ Known
6.2	Tests of Significance
6.3	Use and Abuse of Tests
6.4	Power and Inference as a Decision

Introduction

In this chapter, we show how to use TI calculators to compute confidence intervals and conduct hypothesis tests for the mean μ of a normally distributed population. We begin our study of inference with the unrealistic assumption that the population standard deviation, σ, is known. This assumption will be relaxed in Chapter 7.

6.1 Confidence Intervals with σ Known

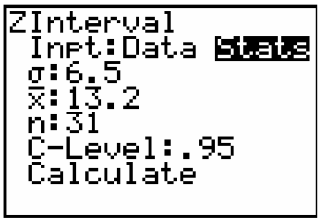
In this section, we show how to compute a confidence interval for the mean of a normal population with known standard deviation σ. To do so, we will use the built-in `ZInterval` feature (item 7) from the `STAT TESTS` menu (item 1 on the [F7] `Ints` menu on a TI-89). The following two exercises demonstrate how to use this feature with summary statistics and with a data set.

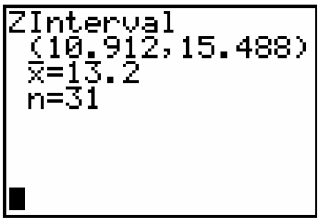


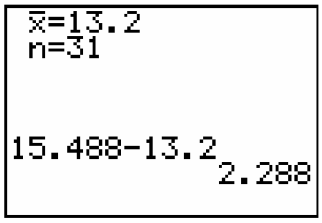
Example 6.1 Bone Turnover. In a study of bone turnover in young women, serum TRAP (a measure of bone resorption) was measured in 31 subjects and the mean was 13.2 Units/liter. Assume that the standard deviation is known to be 6.5 Units/liter. Give the margin of error and find a 95% confidence interval for the mean of all young women represented by this sample.

TI-83/84 Solution. Select the `ZInterval` option, and set `Inpt` to `Stats` by moving the cursor to highlight this option and pressing [ENTER], since we do not have the actual data entered in a list. Enter the given values of 6.5 for σ, 13.2 for $\bar{x}$, and 31 for n. Enter the desired confidence level, then press [ENTER] with `Calculate` highlighted. We obtain a 95% confidence interval of (10.912, 15.488). We normally report one more significant digit than was in the information provided, but check with your instructor for his or her rounding rules. Based on this information, I am 95% confident the mean TRAP level for young women represented by this sample is between 10.92 and 15.49 Units/liter.

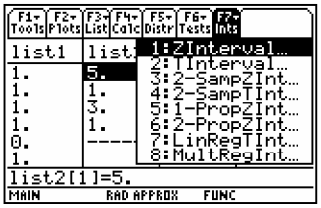
Because the confidence interval is of the form $\bar{x} \pm m$, we can find the margin of error m by subtracting $\bar{x}$ from the right endpoint of the interval: 15.488 – 13.2 = 2.288.

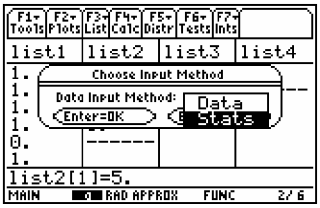


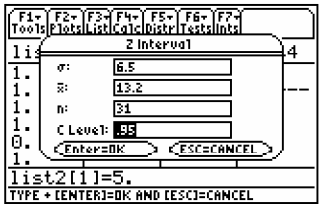




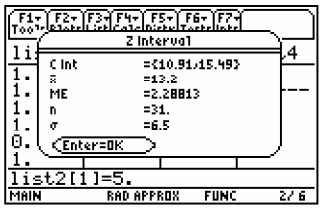
TI-89 Solution. Select `1:ZInterval` from the `Ints` menu. You first see a screen asking for the input method. Press the right arrow to expand the options and move the cursor to the appropriate choice. Here we have Stats, so that should be highlighted. Press [ENTER] to proceed to the next dialog box and enter the summary statistics given above.







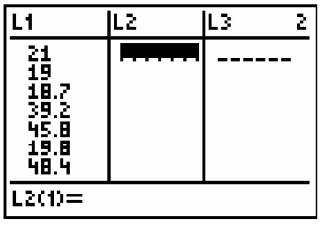
When finished, press ENTER to compute and display the interval. Notice this calculator gives the margin of error explicitly as ME = 2.28813.

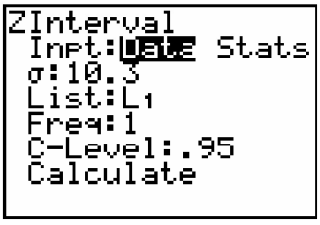


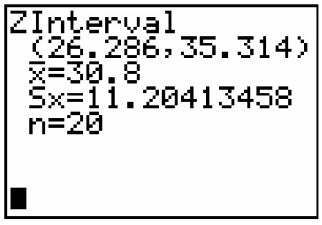
Example 6.2 Fuel Efficiency. Here are the values of the average speed (in mph) for a sample of trials on a vehicle undergoing a fuel efficiency test. Assume that the standard deviation is 10.3 mph. Estimate the mean speed at which the vehicle was driven with 95% confidence.

21.0	19.0	18.7	39.2	45.8	19.8	48.4	21.0	29.1	35.7
31.6	49.0	16.0	34.6	36.3	19.0	43.3	37.5	16.5	34.5

Solution. First, enter the data into a list, say list `L1`. Next, bring up the `ZInterval` screen, set `Inpt` to `Data`, and enter the given value of 10.3 for σ. Set `List` to `L1` with `Freq` 1 (since each value represents one observation), enter the desired confidence level, and press ENTER on `Calculate`. Based on this information, I am 95% confident the mean speed was between 26.29 and 35.31 miles per hour.







Choosing the Sample Size

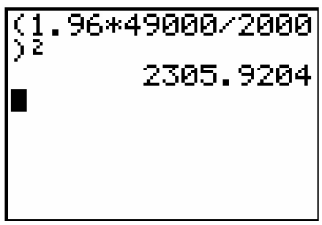
Suppose we want to find the minimum sample size n that will produce a desired margin of error m with a specific level of confidence. To do so, we can use the formula $n \geq \left(\frac{z^* \sigma}{m} \right)^2$, where z^* is the appropriate critical value. We also could use a program that computes the (rounded-up) sample size. To execute the **ZSAMPSZE** program that follows, we simply input the values of the standard deviation σ, the desired margin of error, and the desired confidence level in decimal.

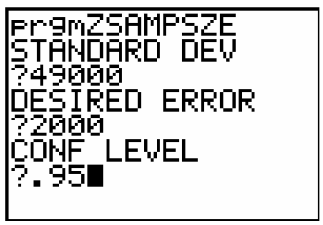
The ZSAMPSZE Program

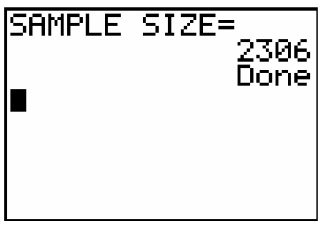
```
PROGRAM:ZSAMPSZE
:Disp "STANDARD DEV."
:Input S
:Disp "DESIRED ERROR"
:Input E
:Disp "CONF. LEVEL"
:Input R
:invNorm((R+1)/2,0,1)STO▶Q
:(Q*S/E)x² STO▶M
:If int(M)=M
:Then
:MSTO▶N
:Else
:int(M+1) STO▶N
:End
:ClrHome
:Disp "SAMPLE SIZE="
:Disp int(N)
```

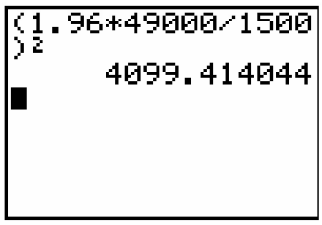
Example 6.3 Student Debt. Suppose we want a margin of error of $2000 with 95% confidence when estimating the mean debt for students completing their undergraduate studies. We have information that the standard deviation is about $49,000. (a) What sample size is required? (b) What sample size would be required to obtain a margin of error of $1500?

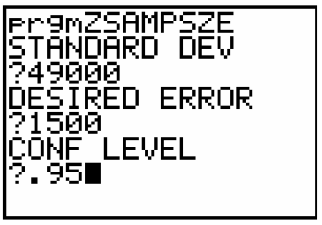
Solution. The critical value for 95% confidence is $z^* = 1.96$. Using this value in the formula $n \geq \left(\frac{1.96 * 49000}{2000} \right)^2$, with σ = 49,000 and m = 2000, we obtain a necessary sample size of n = 2306. The same result is obtained by using the **ZSAMPSIZE** program. Working part (b) similarly with m = 1500, we obtain a required sample size of n = 4100.

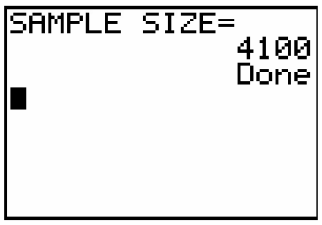






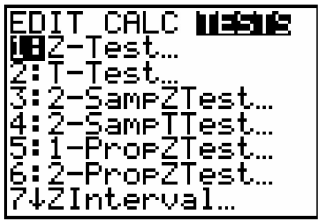






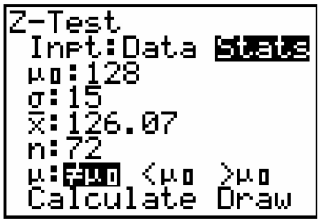
6.2 Tests of Significance

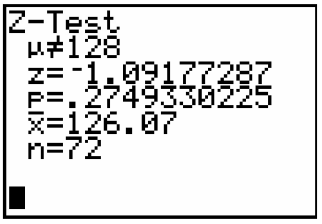
We now show how to perform one-sided and two-sided hypothesis tests about the mean μ of a normally distributed population for which the standard deviation σ is known. To do so, we will use the Z-Test feature (item 1) from the STAT TESTS menu. This menu is found using 2nd F1 =F6 in the Stats/List Editor on a TI-89. We can use this feature to work with either summary statistics or data sets.

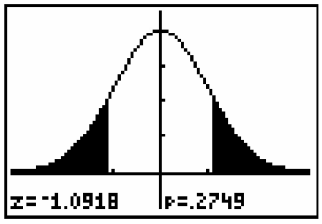


Example 6.4 Executives' Blood Pressure. The mean systolic blood pressure for males 35 to 44 years of age is 128 and the standard deviation is 15. But for a sample of 72 company executives in this age group, the mean systolic blood pressure is $\bar{x}$ = 126.07. Is this evidence that the company's executives in this age group have a different mean systolic blood pressure from the general population?

Solution. To test if the mean is *different* from 128, we use the null hypothesis H_0: μ = 128 with a two-sided alternative H_A: $\mu \neq 128$. Bring up the Z-Test screen and adjust the Inpt to STATS, which allows us to enter the statistics. Enter the values μ_0 = 128, σ = 15, $\bar{x}$ = 126.07, and n = 72. Set the alternative to $\neq \mu_0$ then press ENTER on either Calculate or Draw.





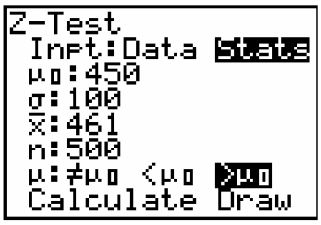


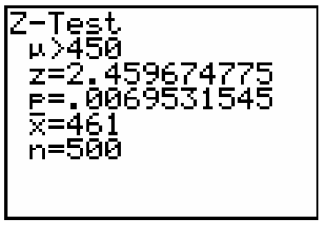
We obtain a z test statistic of –1.0918 and a p-value of 0.2749. For this two-sided test, the p-value comes from the sum of both the right- and left-tail probabilities: $P(z < -1.0918) + P(z > 1.0918)$. If the true mean for all the company's executives in this age group were equal to 128, then there would be a 27.49% chance of obtaining an $\bar{x}$ as

far away as 126.07 with a sample of size 72. This rather high p-value does not give us good evidence to reject the null hypothesis.

Example 6.5 California SATs. An SRS of 500 California high school seniors gave an average SAT mathematics score of $\bar{x} = 461$. Is this good evidence against the claim that the mean for all California seniors is no more than 450? Assuming that $\sigma = 100$ for all such scores, perform the test H_0: $\mu = 450$, H_A: $\mu > 450$. Give the z test statistic and the p-value.

Solution. Bring up the `Z-Test` screen from the `STAT TESTS` menu and check that `Inpt` is `STATS`. Enter the value of $\mu_0 = 450$ and the summary statistics, set the alternative to $> \mu_0$, then scroll down to `Calculate` and press ENTER.



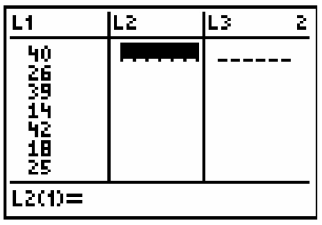


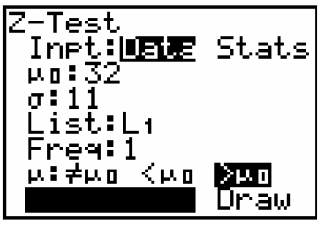
We obtain a z test statistic of 2.46 and a p-value of 0.00695. Because the p-value is so small, we have significant evidence to reject H_0. If the true mean were 450, then there would be only a 0.00695 probability of obtaining a sample mean as high as $\bar{x} = 461$ with an SRS of 500 students.

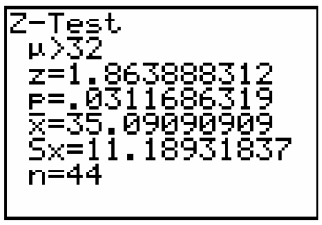
Example 6.6 DRP Scores. The following table gives the DRP scores for a sample of 44 third-grade students in a certain district. It is known that $\sigma = 11$ for all such scores in the district. A researcher believes that the mean score of all third-graders in this district is higher than the national mean of 32. State the appropriate H_0 and H_A, then conduct the test and give the p-value.

40	26	39	14	42	18	25	43	46	27	19
47	19	26	35	34	15	44	40	38	31	46
52	25	35	35	33	29	34	41	49	28	52
47	35	48	22	33	41	51	27	14	54	45

Solution. Here we test H_0: $\mu = 32$ with a one-sided alternative H_A: $\mu > 32$. Enter the data into a list, say list `L1`, then call up the `Z-Test` screen and change `Inpt` to `Data`. Enter the values $\mu_0 = \mathbf{32}$ and $\sigma = \mathbf{11}$, set the list to `L1` with frequencies 1, and set the alternative to $> \mu_0$. Press ENTER on `Calculate` or `Draw`.







We obtain a p-value of 0.0312. If the average of the district were equal to 32, then there would be only a 3.117% chance of a sample group of 44 averaging as high as $\bar{x} = 35.09$. There is evidence to reject H_0 and conclude that the district's average is higher than 32.

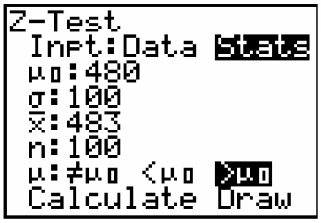
6.3 Use and Abuse of Tests

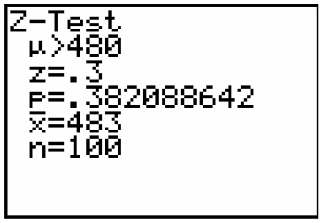
We continue with two more exercises that illustrate how one must be careful in drawing conclusions of significance.

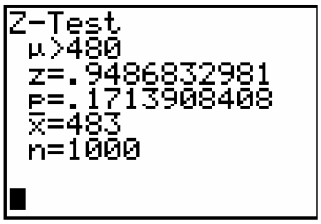
Example 6.7 SAT Coaching. Suppose that SATM scores vary normally with $\sigma = 100$. Calculate the p-value for the test of H_0: $\mu = 480$, H_A: $\mu > 480$ in each of the following situations:

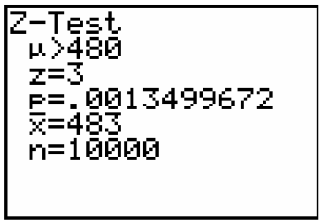
(a) A sample of 100 coached students yielded an average of $\bar{x} = 483$.
(b) A sample of 1000 coached students yielded an average of $\bar{x} = 483$.
(c) A sample of 10,000 coached students yielded an average of $\bar{x} = 483$.

Solution. We adjust the settings in the Z-Test screen from the STAT TESTS menu and calculate, changing n each time. Below are the results using the three different sample sizes. Notice that with increasing sample size, the p-value changes dramatically. We see that the rise in the average score to $\bar{x} = 483$ is significant (p very small) only when the results stem from the very large sample of 10,000 coached students. With the sample of only 100 students, there is 38.2% chance of obtaining a sample mean as high as $\bar{x} = 483$, even if the true mean were still 480.



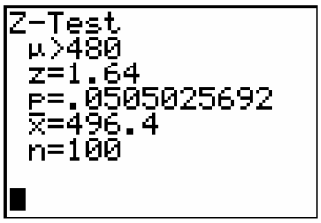


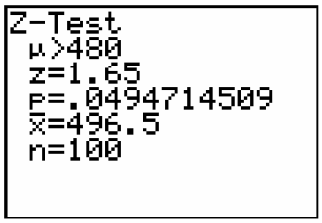




Example 6.8 More SAT Coaching. For the same hypothesis test as in Example 6.7 above, consider the sample mean of 100 coached students. (a) Is $\bar{x} = 496.4$ significant at the 5% level? (b) Is $\bar{x} = 496.5$ significant at the 5% level?

Solution: we perform the `Z-Test` for both values of $\bar{x}$:





In the first case, $p = 0.0505 > 0.05$; so the value of $\bar{x} = 496.4$ is not significant at the 5% level. However, in the second case, $p = 0.04947 < 0.05$; so the value of $\bar{x} = 496.5$ is significant at the 5% level. However, for SATM scores, there is no real "significant" difference between means of 496.4 and 496.5.

6.4 Power and Inference as a Decision

Power is the probability of correctly rejecting a null hypothesis. It is a function of the significance level of the test (with larger α we will correctly reject H_0 more often, but will also wrongly reject it as well), the sample size n, and the distance between H_0 and the true value. We conclude this chapter with some examples on computing the power against an alternative.

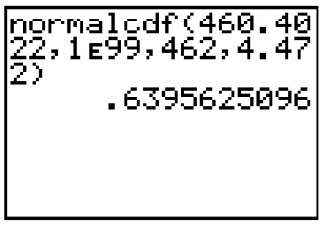
Example 6.9 More California SATs. In Example 6.5 we considered the hypotheses H_0: $\mu = 450$, H_A: $\mu > 450$. A sample of size $n = 500$ is taken from a normal population having $\sigma = 100$. If we performed our test at the 1% level of significance, find the power of this test against the alternative $\mu = 462$.

Solution. We first find the rejection region of the test at the 1% level of significance. Because the alternative is the one-sided right tail, we wish the right-tail probability under the standard normal curve to be 0.01. This probability occurs at $z^* = 2.326$. So we reject H_0 if the z test statistic is more than 2.326. That is, we reject if

$$\frac{\bar{x} - 450}{100/\sqrt{500}} > 2.326$$

or equivalently if $\bar{x} > 450 + 2.326 * 100/\sqrt{500} = 460.4022$. Now we must find the probability that $\bar{x}$ is greater than 460.4022, given that the alternative $\mu = 462$ is true. Given that $\mu = 462$, then $\bar{x} \sim N(462, 100/\sqrt{500} = 4.472)$, and we must compute

$P(\bar{x} > 460.4022)$. To do so, we use the command `normalcdf(460.4022,1E99,462,4.472)` and find that the power against the alternative $\mu = 462$ is about 0.64.

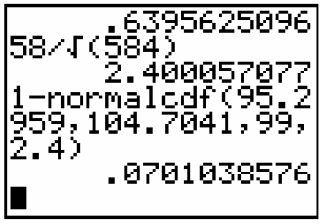


Example 6.10 Two-sided Power. (a) An SRS of size 584 is taken from a population having $\sigma = 58$ to test the hypothesis H_0: $\mu = 100$ versus a two-sided alternative at the 5% level of significance. Find the power against the alternative $\mu = 99$.

Solution. Again, we first must find the rejection regions. For a two-sided alternative at the 5% level of significance, we allow 2.5% at each tail. Thus, we reject if the z test statistic is beyond $\pm$ 1.96. That is, we reject if

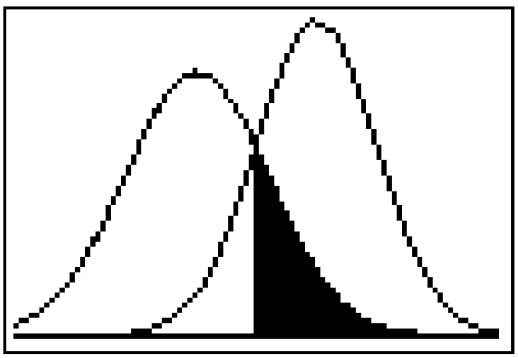
$$\frac{\bar{x}-100}{58/\sqrt{584}} < -1.96 \quad \text{or} \quad \frac{\bar{x}-100}{58/\sqrt{584}} > 1.96$$

Equivalently, we reject if $\bar{x} < 95.2959$ or if $\bar{x} > 104.7041$. Now assuming that $\mu = 99$, then $\bar{x} \sim (99, 58/\sqrt{584} = 2.400)$. We now must compute $P(\bar{x} < 95.2959) + P(\bar{x} > 104.7041)$, which is equivalent to $1 - P(95.2959 < \bar{x} < 104.7041) =$ `1-normalcdf(95.2959,104.7041,99,2.400)`. With this command, we see that the power against the alternative $\mu = 99$ is about 0.07. This low power is to be expected, since the difference between 100 and 99 is small.



CHAPTER

7



Inference for Distributions

7.1	Inference for the Mean of a Population
7.2	Comparing Two Means
7.3	Optional Topics in Comparing Distributions

Introduction

In this chapter, we demonstrate the various t procedures that are used for confidence intervals and significance tests about the mean of a normal population for which the standard deviation is unknown. We also consider comparing means from two independent samples and paired samples.

7.1 Inference for the Mean of a Population

We begin with a short program that allows us to find a critical value t^* after specifying the degrees of freedom and confidence level. This is a built-in function on TI-84 and TI-89 calculators. TI-83s need this program (or you could use Table C). This program asks for degrees of freedom rather than sample size, since t distributions are used in inference for more than just a single sample.

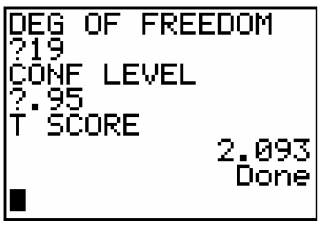
The TSCORE Program

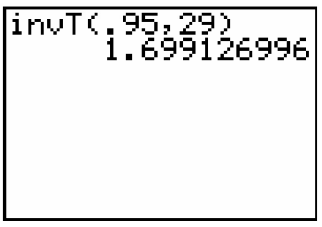
PROGRAM:TSCORE	:tcdf(0,X,M)STO▸Y1
:Disp "DEG. OF FREEDOM"	:solve(Y1-R/2,X,2) STO▸Q
:Input M	:Disp "T SCORE"
:Disp "CONF. LEVEL"	:Disp round(Q,3)
:Input R	

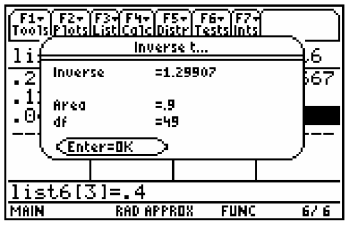
Example 7.1 Finding t^* Find the critical values t^* for confidence intervals for the mean in the following cases:

(a) A 95% confidence interval based on $n = 20$ observations
(b) A 90% confidence interval from an SRS of 30 observations
(c) An 80% confidence interval from a sample of size 50

Solution. The confidence intervals are based on t distributions with $n - 1$ degrees of freedom, since we have single samples. So we need 19 degrees of freedom for part (a), 29 degrees of freedom for part (b), and 49 degrees of freedom for part (c). Below are the outputs of the **TSCORE** program and from the built-in functions. These (like invNorm) use area to the left of the point desired, so for a 90% confidence interval there is .95 to the left of t^* and for 80% confidence there is 90% to the left of t^*.





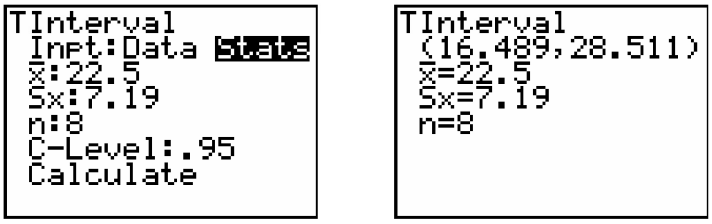


One-Sample *t* Confidence Interval

We now examine confidence intervals for one mean for which we will use the TInterval feature (item 8) from the STAT TESTS menu ([F7] Ints on a TI-89). As with the ZInterval feature that we used in Chapter 6, we can enter the summary statistics or use data in a list.

Example 7.2 Vitamin C in Corn. The amount of vitamin C in a factory's production of corn soy blend (CSB) is measured from eight samples giving $\bar{x}$ = 22.50 (mg/100 g) and s = 7.19. Find a 95% confidence interval for the mean vitamin C content of the CSB produced during this run.

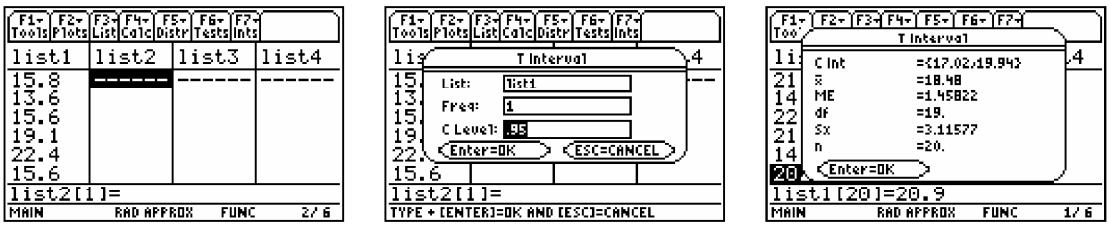
Solution. Call up the TInterval feature from the STAT TESTS menu. Set Inpt to Stats, then enter the values of $\bar{x}$, Sx, n, C-Level, and press [ENTER] on Calculate. We obtain the interval 16.489 to 28.511.



Example 7.3 Fuel Efficiency. Here are the values of the fuel efficiency in mpg for a sample of trials on a vehicle undergoing testing. Find the mean, the standard deviation, the standard error, the margin of error for a 95% confidence interval, and give a 95% confidence interval for the mean mpg of this vehicle.

15.8	13.6	15.6	19.1	22.4	15.6	22.5	17.2	19.4	22.6
19.4	18.0	14.6	18.7	21.0	14.8	22.6	21.5	14.3	20.9

Solution. First, enter the data into a list, say list L1 (or list1 on a TI-89). Next, bring up the TInterval screen, set Inpt to Data, set the List to L1 with frequencies 1, enter the desired confidence level, and press [ENTER] on Calculate. The sample mean, sample deviation, and confidence interval are all displayed.



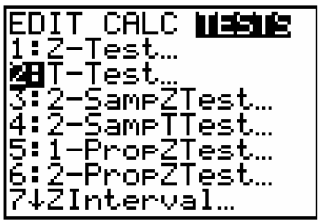
Because the confidence interval is of the form $\bar{x} \pm m$, we can find the margin of error m by subtracting $\bar{x}$ from the right endpoint of the interval: $m = 19.938 - 18.48 = 1.458$. The TI-89 gives the margin of error explicitly. The standard error is given by $s/\sqrt{n} = 3.115766358/\sqrt{20} = 0.6967$.

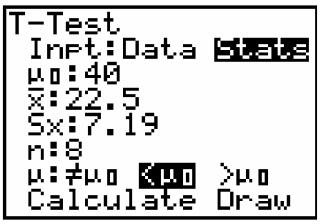
One-Sample *t* test

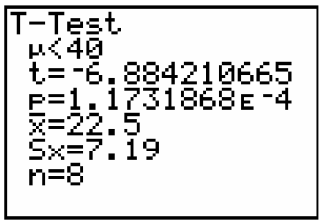
We now perform some significance tests about the mean using the `T-Test` feature (item 2) from the `STAT TESTS` menu.

Example 7.4 More Vitamin C in Corn. Using the vitamin C data of $n = 8$, $\bar{x} = 22.50$, and $s = 7.19$ from Example 7.2 above, test the hypothesis H_0: $\mu = 40$ versus the alternative H_A: $\mu < 40$.

Solution. Bring up the `T-Test` screen from the `STAT TESTS` menu and adjust `Inpt` to `STATS`. Enter the value of $\mu_0 = 40$ and the summary statistics, set the alternative to $< \mu_0$, then scroll down to `Calculate` and press ENTER.





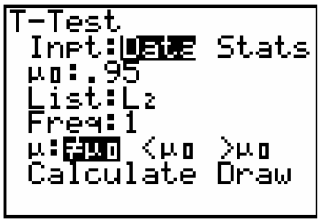


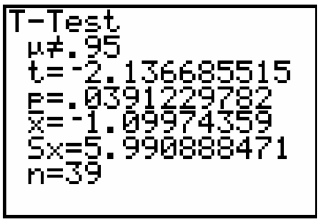
We obtain a t test statistic of –6.88 and a p-value of 0.0001. Because the p-value is so small, we have significant evidence to reject H_0. If the true mean were 40, a sample mean of 22.5 or lower with a sample of size 8 should only happen about 1 in 10,000 times.

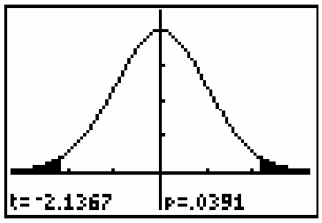
Example 7.5 Stock Returns. An investor sued his broker and brokerage firm because lack of diversification in his portfolio led to poor performance. The following table gives the monthly percentage rates of return for the months in which the account was managed by the broker. For these months, the average of the Standard & Poor's 500-Stock Index was 0.95%. Are these returns compatible with the S&P 500 average? Use the data to test the hypothesis H_0: $\mu = 0.95$ versus the alternative H_A: $\mu \neq 0.95$.

–8.36	1.63	–2.27	–2.93	–2.70	–2.93	–9.14	–2.64
6.82	–2.35	–3.58	6.13	7.00	–15.25	–8.66	–1.03
–9.16	–1.25	–1.22	–10.27	–5.11	–0.80	–1.44	1.28
–0.65	4.34	12.22	–7.21	–0.09	7.34	5.04	–7.24
–2.14	–1.01	–1.41	12.03	–2.56	4.33	2.35	

Solution. Enter the data into a list, say list L2, then bring up the T-Test screen and set Inpt to Data. Enter the value of $\mu_0 = .95$, set the list to L2 with frequency 1, and set the alternative to $\neq \mu_0$. Then press ENTER on Calculate or Draw.







We obtain a *p*-value of 0.0391. If the average return were equal to 0.95%, then there would be only a 3.91% chance of a sample of 39 months averaging as far or farther away from that value as –1.1%. There is sufficient evidence to reject H_0 and conclude that the mean monthly return differs from (was worse than) 0.95%.

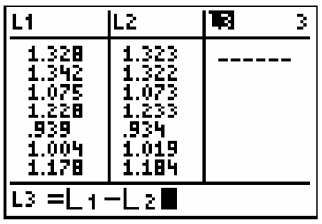
Matched Pair *t* Procedure

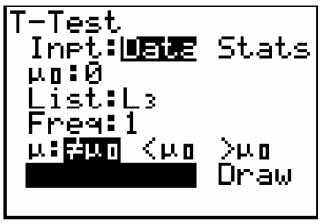
Example 7.6 Are the Technicians Consistent? Two operators of X-ray machinery measured the same eight subjects for total body bone mineral content. We want these two operators to have consistent results when dealing with this test. Here are the results in grams:

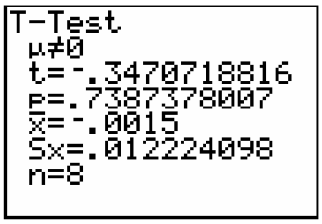
	Subject							
Operator	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	1.328	1.342	1.075	1.228	0.939	1.004	1.178	1.286
2	1.323	1.322	1.073	1.233	0.934	1.019	1.184	1.304

Use a significance test to examine the null hypothesis that the two operators have the same mean. Use a 95% confidence interval to provide a range of differences that are compatible with these data.

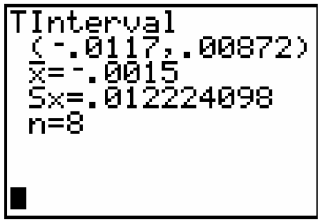
Solution. We consider the average μ_D of the *difference* of the measurements between the operators. First, enter the measurements of Operator 1 into list L1 and the measurements of Operator 2 into list L2. Next, highlight list name L3 and enter the command L1-L2 and press ENTER to store the differences. Then, use a T-Test on list L3 to test H_0: $\mu_D = 0$ versus the alternative H_A: $\mu_D \neq 0$.







We obtain a p-value of 0.7387 from a test statistic of $t = -0.347$. Due to the high p-value, we can say that there is not a significant average difference. If μ_D were equal to 0, then there would be a 73.87% chance of having an average difference as far away as $\bar{d} = -0.0015$ with a random sample of eight subjects. Next, use the `TInterval` on list `L3` to find a 95% confidence interval for the average difference. Because the interval (–0.0117, 0.00872) contains 0, we have further evidence that the operators could have the same mean.



The Sign Test

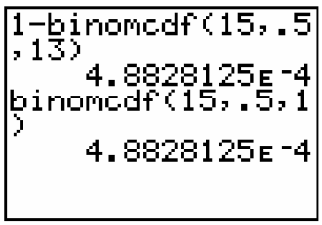
The sign test is an easy way to perform a hypothesis test about the median of the distribution if the data are not normal. It is based on the binomial distribution. If the null hypothesis is true there should be a 50% chance of observing a response either more than or less than the claimed median.

Example 7.7 The Full Moon and Behavior. A study of dementia patients in nursing homes recorded various types of disruptive behavior every day for 12 weeks. Days were classified as moon days if they were in a three-day period centered at the day of the full moon. For each patient the average number of disruptive behaviors for moon days and for all other days was tracked. Out of 15 patients, 14 had more aggressive behavior on moon days than on other days. Use the sign test on the hypothesis of "no moon effect."

Solution. Because so many patients had a change in behavior, we shall test the hypothesis H_0: $p = 0.50$ with the alternative H_A: $p > 0.50$.

We must compute the probability of there being as many as 14 changes with a $B(15,0.50)$ distribution. Equivalently, we can compute the probability of there being as few as one nonchange. Thus, we could compute either $P(B \geq 14) = 1 - P(B \leq 13)$ or $P(B \leq 1)$. To do so, we use the built-in `binomcdf(` command from the `DISTR` menu.

After entering either `1-binomcdf(15,.5,13)` or `binomcdf(15,.5,1)`, we obtain the very low p-value of 0.000488. If there were no moon effect, then there would be almost no chance of having as many as 14 out of 15 showing a change. Therefore, we can reject H_0 in favor of the alternative that the moon generally causes more aggressive behavior.



7.2 Comparing Two Means

We next consider confidence intervals and significance tests for the difference of means $\mu_1 - \mu_2$ given two normal populations that have unknown standard deviations. The results are based on independent random samples of sizes n_1 and n_2. For the most accurate results, we can use the `2-SampTInt` and `2-SampTTest` features from the `STAT TESTS` menu.

These features require that we specify whether or not we wish to use the pooled sample variance $s_p{}^2$. We should specify "Yes" only when we assume that the two populations have the *same* (unknown) variance. In this case, the critical values t^* are obtained from the $t(n_1 + n_2 - 2)$ distribution and the standard error is $s\sqrt{1/n_1 + 1/n_2}$. When we specify "No" for the pooled variance, then the standard error is $\sqrt{s_1^2/n_1 + s_2^2/n_2}$ and the degrees of freedom r are given by

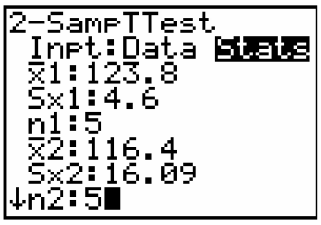
$$r = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

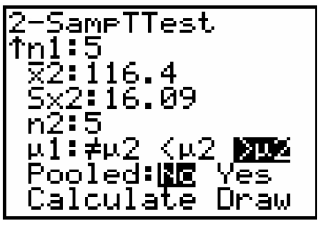
But if the true population standard deviations σ_1 and σ_2 are known, then we should use the `2-SampZInt` and `2-SampZTest` features for our calculations.

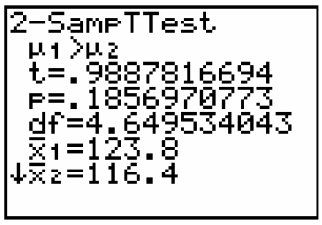
Example 7.8 Does Polyester Decay? How quickly do synthetic fabrics decay in landfills? A researcher buried polyester strips for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and a good indicator of decay. One set of strips was dug up after 2 weeks and the other was dug up after 16 weeks. We suspect that decay increases over time. The results are given in the table below. Test the hypothesis H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 > \mu_2$, then give a 95% confidence interval for $\mu_1 - \mu_2$.

Group	n	$\bar{x}$	s
2 weeks	5	123.80	4.60
16 weeks	5	116.40	16.09

Solution. Use the `2-SampTTest` feature from the `STAT TESTS` menu, and set `Inpt` to `Stats`. Enter the given statistics, set the alternative, and enter `No` for `Pooled`. Then press `ENTER` on `Calculate`.

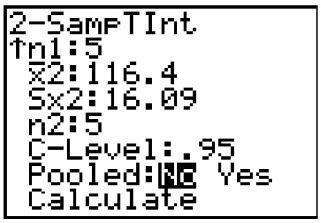


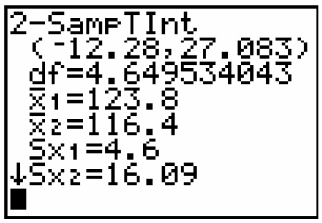




We obtain a p-value of 0.1857 from a test statistic of .9888 with 4.6495 degrees of freedom. If the true mean decay rates were equal, then there would be a good chance of our means being as different as they are (the difference is 7.4) with samples of these sizes. We therefore fail to reject H_0 and conclude that there is not a significant difference in decay of polyester with the extra 14 weeks.

To calculate a confidence interval for $\mu_1 - \mu_2$, we use `2-SampTInt`. Our data is still there, so one can simply move the arrow down to enter the desired confidence level and then calculate. We obtain the interval (–12.28, 27.08). These were small samples, which is one reason for the wide interval. We also note that 0 is included in the interval, which supports our conclusion of no significant increase in polyester decay with the additional 14 weeks of burial.





Example 7.9 College Study Habits. The Survey of Study Habits and Attitudes was given to first-year students at a private college. The tables below show a random sample of the scores.

Women's scores

154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

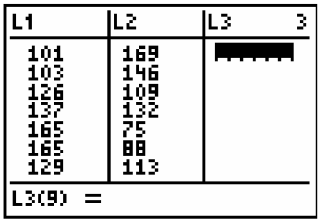
Men's scores

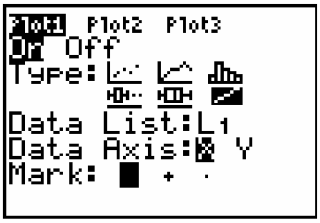
108	140	114	91	180	115	126	92	169	146
109	132	75	88	113	151	70	115	187	104

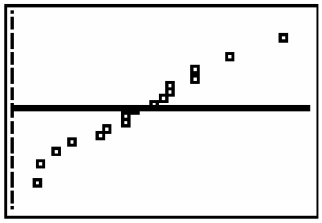
(a) Examine each sample graphically to determine if the use of a t procedure is acceptable.

(b) Test the supposition that the mean score for all men is lower than the mean score for all women among first-year students at this college.
(c) Give a 90% confidence interval for the mean difference between the SSHA scores of male and female first-year students at this college.

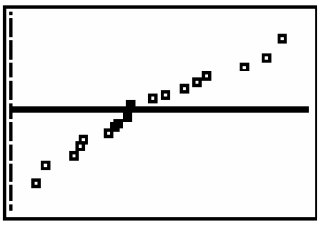
Solution. (a) We shall make normal quantile plots of these data. Enter the women's scores into list L1 and the men's scores into list L2. In the STAT PLOT screen, adjust the Type settings for both Plot1 to the last type for a normal quantile plot, then graph each list separately. Press ZOOM 9 to display the plots. The resulting plots appear close enough to linear to warrant use of *t* procedures.





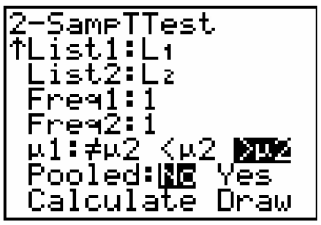


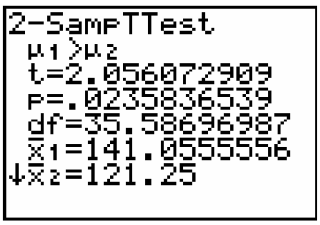
Women

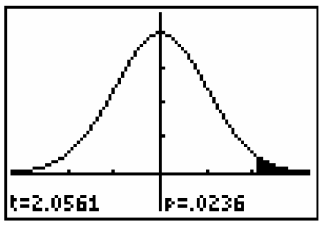


Men

(b) Next, let μ_1 be the mean SSHA score among all first-year women and let μ_2 be the mean score among all first-year men. We shall test the hypothesis H_0: $\mu_1 = \mu_2$ versus the alternative H_A: $\mu_1 > \mu_2$. On the 2-SampTTest screen, set Inpt to Data, enter the desired lists L1 and L2, set the alternative to $> \mu_2$, enter No for Pooled, and press ENTER on Calculate or Draw.



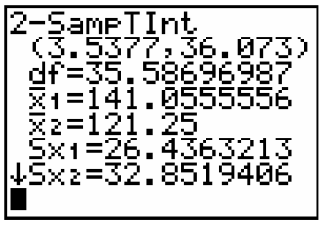




We obtain a *p*-value of 0.02358. If the true means were equal, then there would be only a 2.358% chance of $\bar{x}_1$ being so much larger than $\bar{x}_2$ with samples of these sizes. The

relatively low *p*-value gives us evidence to reject H_0 and conclude that $\mu_1 > \mu_2$. That is, the mean score for all men is lower than the mean score for all women among first-year students at this college.

(c) Adjust the settings in the 2-SampTInt screen and calculate. We obtain (3.5377, 36.073). That is, the mean score of female first-year students should be from about 3.5377 points higher to about 36.073 points higher than the mean score of male first-year students at this college.



Pooled Two-Sample *t* Procedures

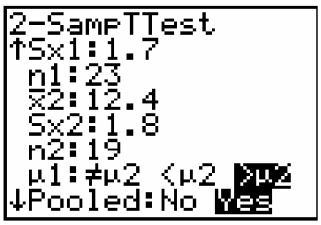
Example 7.10 Infants' Hemoglobin. A study of iron deficiency in infants compared samples of infants following different feeding regimens. Here are the summary results on hemoglobin levels at 12 months of age for two samples of infants:

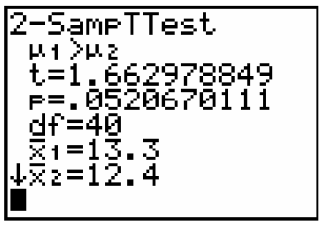
Group	*n*	$\bar{x}$	*s*
Breast-fed	23	13.3	1.7
Formula	19	12.4	1.8

(a) Is there significant evidence that the mean hemoglobin level is higher among breast-fed babies? State H_0 and H_A, and carry out a *t* test.
(b) Give a 95% confidence interval for the difference in mean hemoglobin levels between the two populations of infants.

Solution. Let μ_1 be the mean hemoglobin level for all breast-fed babies and let μ_2 be the mean level for all formula-fed babies. Because the sample deviations are so close, it appears that true standard deviations among the two groups could be equal; thus, we may use the pooled two-sample *t* procedures. For part (a), we will test H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 > \mu_2$.

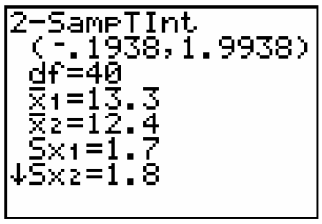
Call up the 2-SampTTest feature, and set Inpt to Stats. Enter the given statistics, set the alternative, enter Yes for Pooled, and calculate.





With a p-value of 0.052, we conclude that there is not statistical evidence (at the 5% level) to reject H_0. If the true means were equal, then there is greater than a 5% chance of $\bar{x}_1$ being 0.9 higher than $\bar{x}_2$ with samples of these sizes.

(b) Next, calculate a 95% confidence interval with the 2-SampTInt feature set to Yes on Pooled. We see that $-0.1938 < \mu_1 - \mu_2 < 1.9938$. That is, the mean hemoglobin level of breast-fed babies could be from 0.1938 lower to 1.9938 higher than the mean level of formula-fed babies. Since 0 is in the interval, we are further convinced that there is not a significant difference between breast-fed and formula-fed babies' hemoglobin levels, on average.



7.3 Optional Topics in Comparing Distributions

We now demonstrate a test for determining whether or not two normal populations have the same variance. If so, then we would be justified in using the pooled two-sample t procedures for confidence intervals and significance tests about the difference in means. For the test, we will need the 2-SampFTest feature from the STAT TESTS menu.

One word of caution about this test is in order, however. This is *extremely* sensitive to any departures from normality, and the lack of robustness does not improve substantially with larger sample sizes. As such, it should be used with wariness; it is difficult to tell whether a significant p-value is due to a difference in the standard deviations, or to a lack of normality.

The *F* Ratio Test

Example 7.11 More on SSHA. Consider again the data from Example 7.9 regarding the SSHA scores of first-year students at a private college. Test whether the women's scores are less variable than the men's.

Women's scores

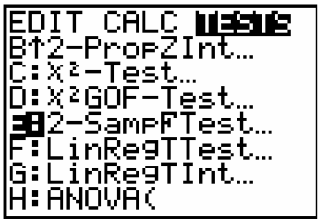
154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

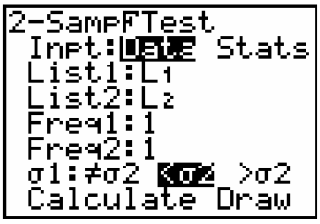
Men's scores

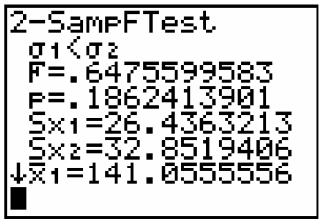
108	140	114	91	180	115	126	92	169	146
109	132	75	88	113	151	70	115	187	104

Solution. Let σ_1 be the standard deviation of all women's scores and let σ_2 be the standard deviation for all men's scores. We shall test H_0: $\sigma_1 = \sigma_2$ versus H_A: $\sigma_1 < \sigma_2$. To do so, first enter the data sets into lists, say L1 and L2. Next, bring up the

2-SampFTest screen from the STAT TESTS menu, set Inpt to Data, enter the appropriate lists names and alternative, and calculate.





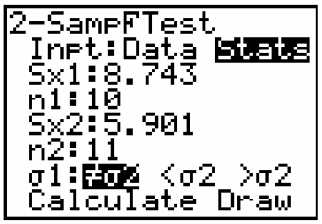


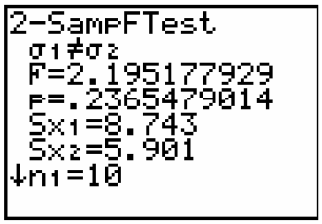
With a *p*-value of 0.18624, we do not have strong evidence to reject H_0. If σ_1 were equal to σ_2, then there would be an 18.6% chance of the women's sample deviation of Sx1 = 26.4363 being so much lower than the men's sample deviation of Sx2 = 32.8519. Thus, we cannot assert strongly that the women's scores are less variable.

Example 7.12 Calcium and Blood Pressure. Here are the summary statistics for the drop in blood pressure from two sample groups of patients undergoing treatment. Test to see if the groups in general have the same standard deviation.

Group	*n*	$\bar{x}$	*s*
Calcium	10	5.000	8.743
Placebo	11	–0.273	5.901

Solution. Let σ_1 be the standard deviation of all possible patients in the calcium group, and let σ_2 be the standard deviation of all possible patients in the placebo group. We will test the hypothesis H_0: $\sigma_1 = \sigma_2$ versus H_A: $\sigma_1 \neq \sigma_2$. Bring up the 2-SampFTest screen and set Inpt to Stats. Enter the summary statistics and alternative, then calculate.

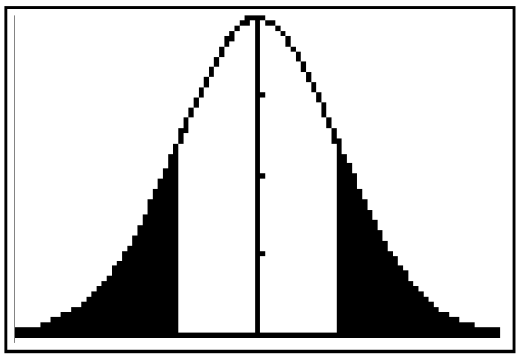




We obtain an *F*-statistic of 2.195 and a *p*-value of 0.2365. If σ_1 were equal to σ_2, then there would be about a 23.6% chance of Sx1 and Sx2 being so far apart with samples of these sizes. This evidence may not be significant enough to reject H_0 in favor of the alternative.

CHAPTER

8



Inference for Proportions

8.1	Inference for a Single Proportion
8.2	Comparing Two Proportions

Introduction

In this chapter, we discuss how to find confidence intervals and to conduct hypothesis tests for a single proportion and for the difference in two population proportions. We will also show how to adjust the counts for confidence intervals in which there is a small number of successes or failures.

8.1 Inference for a Single Proportion

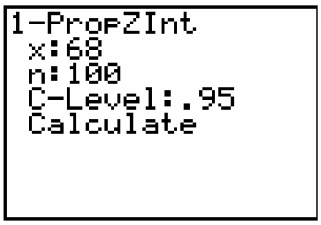
Both the large-sample and plus-four level C confidence intervals can be calculated using the `1-PropZInt` feature from the `STAT TESTS` menu. Significance tests can be worked using the `1-PropZTest`.

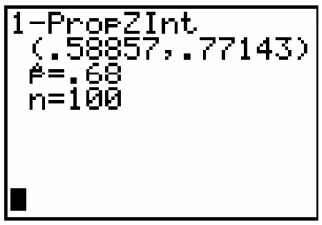


A Large-Sample Confidence Interval

Example 8.1 Stress in Restaurant Workers. In a restaurant worker survey, 68 of a sample of 100 employees agreed that work stress had a negative impact on their personal lives. Find a 95% confidence interval for the true proportion of restaurant employees who agree.

Solution. Bring up the `1-PropZInt` screen, enter 68 for `x`, enter 100 for `n`, and enter .95 for `C-Level`. Then press `ENTER` on `Calculate` to obtain a 95% confidence interval of (0.58857, 0.77143). Based on this sample, with 95% confidence, between 58.9% and 77.1% of restaurant workers will agree that work stress has a negative impact on their personal lives.



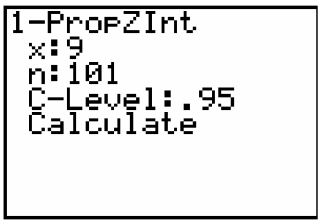


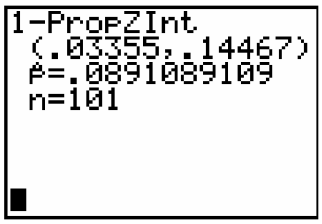
A Plus-Four Confidence Interval

Example 8.2 Blind Medical Trials. Many medical trials randomly assign patients to either an active treatment or a placebo. The trials are supposed to be double-blind, but sometimes patients can tell whether or not they are getting the active treatment. Reports of medical research usually ignore the problem. Investigators looked at a random sample of 97 articles and found that only 7 discussed their success in blinding the trial. What proportion of all such studies discuss the success of blinding in their trials? Give a 95% confidence interval estimate.

Because we have a small number of studies that discussed the success of blinding in the trial (less than 10), we use the "plus-four" method: we simply add 4 to the number of studies (it becomes 101) and 2 to the number of "successes" (we'll consider that 9 discussed the success in blinding the trial). We use `1-PropZInt` as before, with the adjusted counts.

Solution. In the `1-PropZInt` screen, enter 9 for `x`, which is 2 more than the actual number and enter 101 for `n`, which is 4 more than the actual sample size. Enter the desired `C-level` (95%) and press [ENTER] on `Calculate` to obtain the plus-four estimate $\tilde{p} = 0.089$ and the confidence interval.





Based on this sample, we are 95% confident that between 3.4% and 14.5% of published medical studies will discuss the success of blinding in their trials.

Choosing a Sample Size

As with confidence intervals for the mean, we often would like to know in advance what sample size would provide a certain maximum margin of error m with a certain level of confidence. The required sample size n satisfies $n \geq \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$, where z^* is the appropriate critical value depending on the level of confidence and p^* is a guessed value of the true proportion p. If $p^* = 0.50$, then the resulting sample size insures that the margin of error is no more than m, regardless of the true value of p.

While the formula is fairly easy to use, we supply a simple program **PSAMPSZE** that displays the required sample size, rounded up to the nearest integer, after one enters the desired error m, the confidence level, and the guess p^*.

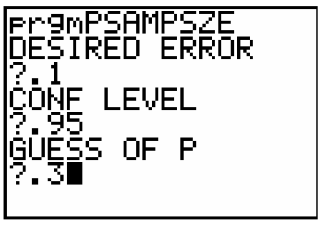
The PSAMPSZE Program

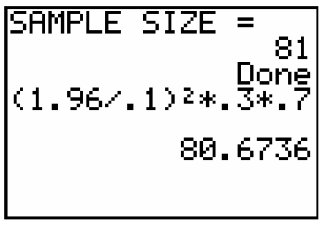
```
PROGRAM:PSAMPSZE
:Disp "DESIRED ERROR"
:Input E
:Disp "CONF. LEVEL"
:Input R
:Disp "GUESS OF P"
:Input P
:invNorm((R+1)/2,0,1)STO▶Q
:(Q/E)x²*P(1-P) STO▶M
:If int(M)=M
:Then
:MSTO▶N
:Else
:int(M+1) STO▶N
:End
:ClrHome
:Disp "SAMPLE SIZE="
:Disp int(N)
```

Example 8.3 Alcohol Awareness — Find a Sample Size. Among students who completed an alcohol awareness program, you want to estimate the proportion who state that their behavior towards alcohol has changed since the program. Using the guessed

value of $p^* = 0.30$ from previous surveys, find the sample size required to obtain a 95% confidence interval with a maximum margin of error of $m = 0.10$.

Solution. Executing the **PSAMPSZE** program, we find that a sample of size 81 would be required. This value also can be obtained by $n = (1.96/.1)^2 * .3 * .7 = 80.6736$ which we round up (as always for sample sizes) to 81.



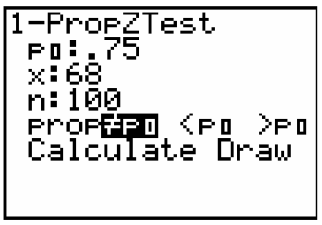


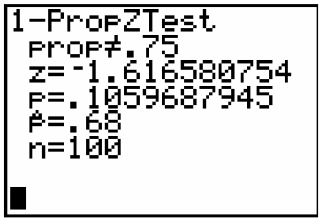
Significance Tests

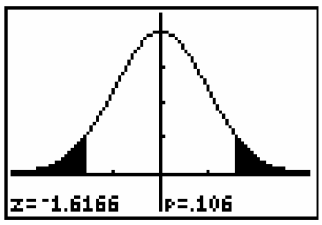
We now show how to conduct hypothesis tests for a single population proportion p using the 1-PropZTest feature (item 5) in the STAT TESTS menu.

Example 8.4 More Stress in Restaurant Workers. In the restaurant worker survey, 68 of a sample of 100 employees agreed that work stress had a negative impact on their personal lives. Let p be the true proportion of restaurant employees who agree. Test the hypothesis H_0: $p = 0.75$ versus H_A: $p \neq 0.75$.

Solution. Enter the data and alternative into the 1-PropZTest screen, then press ENTER on Calculate or Draw. We obtain a (two-sided) p-value of 0.106 from a z-statistic of –1.61658. If p were equal to 0.75, then there would be a 10.6% chance of obtaining $\hat{p}$ as far away as 0.68 with a sample of size 100. We therefore believe the proportion of restaurant workers who would agree that work stress has a negative impact on their personal lives is not significantly different from 75%.

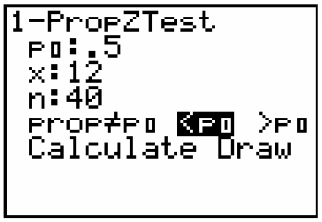


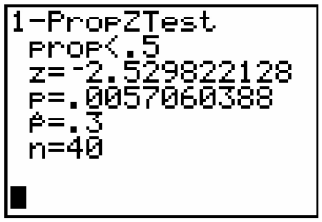




Example 8.5 Who Likes Instant? In a taste test of instant versus fresh-brewed coffee, only 12 out of 40 subjects preferred the instant coffee. Let p be the true probability that a random person prefers the instant coffee. Test the claim H_0: $p = 0.50$ versus H_A: $p < 0.50$ at the 5% level of significance.

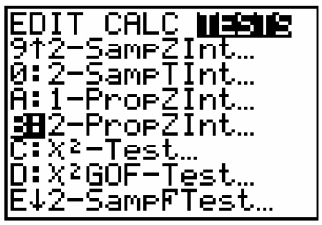
Solution. Enter the data and alternative into the 1-PropZTest screen and calculate. We obtain a test statistic of –2.53 and a p-value of 0.0057. If p were 0.50, then there would be only a 0.0057 probability of $\hat{p}$ being as low as 0.3 with 40 subjects. There is strong evidence to reject H_0, and conclude that those who prefer instant coffee are a minority of the population.





8.2 Comparing Two Proportions

We now demonstrate confidence intervals and significance tests for the difference of two population proportions (p_1 and p_2). These calculations are done using the 2-PropZInt and 2-PropZTest (item 5) features from the STAT TESTS menu.

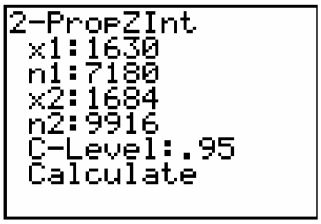


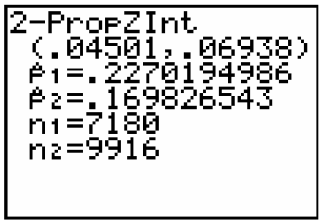
A Large-Sample Confidence Interval for Difference of Proportions

Example 8.6 Binge Drinking on Campus. The table below gives the sample sizes and numbers of men and women who responded "Yes" to being frequent binge drinkers in a survey of college students. Find a 95% confidence interval for the difference between the proportions of men and women who are frequent binge drinkers.

Population	n	X
Men	7180	1630
Women	9916	1684

Solution. Call the men group 1 and the women group 2. In the 2-PropZInt screen, enter **1630** for x1, **7180** for n1, **1684** for x2, and **9916** for n2. Set the confidence level to **.95** and press ENTER on Calculate.





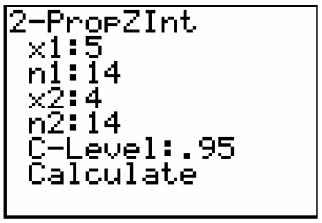
We obtain a confidence interval of (0.045, 0.069). That is, the proportion of male binge drinkers is from 4.5 percentage points higher to 6.9 percentage points higher than the proportion of female binge drinkers, with 95% confidence.

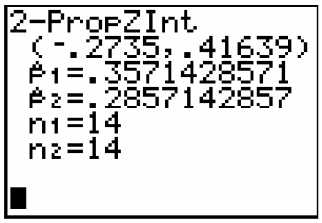
A Plus-Four Confidence Interval for Difference of Proportions

Example 8.7 Gender Differences. In studies that look for a difference between genders, a major concern is whether or not apparent differences are due to other variables that are associated with gender. A study of 12 boys and 12 girls found that 4 of the boys and 3 of the girls had a Tanner score (a measure of sexual maturity) of 4 or 5 (a high level of sexual maturity). Find a plus-four 95% confidence interval for the difference in proportions of all boys and all girls who would score a 4 or 5.

Solution. We can still use the `2-PropZInt` feature to find a plus-four confidence interval for $p_1 - p_2$. But for `x1` and `x2`, enter 1 more than the actual number of positive responses (we split the two successes added in the one sample case). For `n1` and `n2`, enter 2 more than the actual sample sizes (splitting the added four trials). Here, enter **5** for `x1`, **4** for `x2`, and enter **14** for both `n1` and `n2`. Then, set the confidence level to **.95** and calculate.

We see that $-0.2735 < p_1 - p_2 < 0.4164$. That is, the true proportion of boys who score 4 or 5 is from 27.35 percentage points lower to 41.64 percentage points higher than the true proportion of girls who score 4 or 5. Since 0 is included in this interval, there may be no difference in the population proportions of Tanner scores.





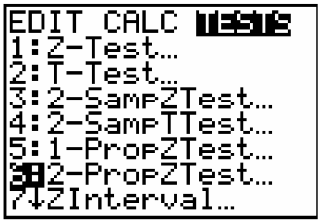
Significance Tests for Difference of Proportions

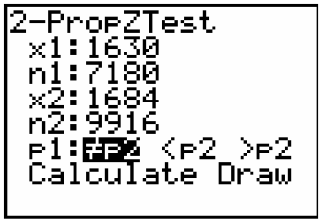
We now show how to conduct hypothesis tests about $p_1 - p_2$ using the `2-PropZTest` feature from the `STAT TESTS` menu.

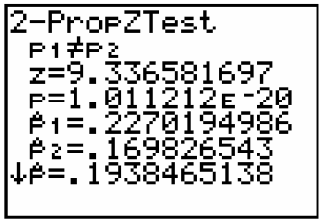
Example 8.11 More Binge Drinking. The table below again gives the sample sizes and numbers of men and women who responded "Yes" to being frequent binge drinkers in a survey of college students. Does the data give good evidence that the true proportions of male binge drinkers and female binge drinkers are different?

Population	n	X
Men	7180	1630
Women	9916	1684

Solution. Let p_1 be the true proportion of male students who are frequent binge drinkers, and let p_2 be the true proportion of female students who are frequent binge drinkers. We shall test the hypothesis H_0: $p_1 = p_2$ versus the alternative H_A: $p_1 \neq p_2$. Bring up the `2-PropZTest` screen, enter the actual data and the two-sided alternative, then calculate.





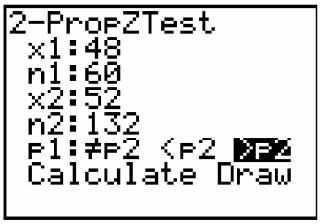


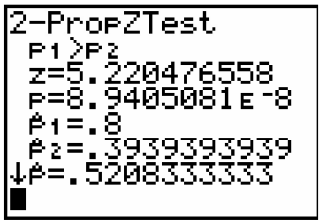
We obtain a p-value of 1×10^{-20} (essentially 0) from a test statistic of $z = 9.33658$. If $p_1 = p_2$ were true, then there would be no chance of obtaining sample proportions as far apart as $\hat{p}_1 = 0.227$ and $\hat{p}_2 = 0.1698$ with samples of these sizes. So, we can reject H_0 and conclude that not only are the genders different in (admitting to) binge drinking, but that males are also more likely to engage in binge drinking (or at least to admit to it). This reinforces our results obtained in Example 8.6; the confidence interval was wholly positive, so there would be no belief that the two proportions should be the same.

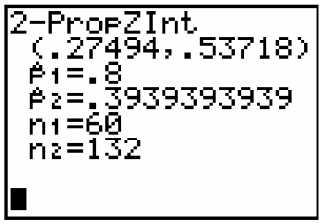
Example 8.12 Juvenile References. The table below gives the results of a gender bias analysis of a textbook. Do the data give evidence that the proportion of juvenile female references is higher than the proportion of juvenile male references? Compare the results with those of a 90% confidence interval.

Gender	n	X (juvenile)
Female	60	48
Male	132	52

Solution. Let p_1 be the true proportion of juvenile female references in all such texts, and let p_2 be the true proportion of juvenile male references. We will test the hypothesis H_0: $p_1 = p_2$ versus H_A: $p_1 > p_2$. Bring up the `2-PropZTest` screen, enter the data, set the alternative to `>p2`, and calculate. Then bring up the `2-PropZInt` screen, and calculate a 90% confidence interval.







We obtain a very low p-value of 8.94×10^{-8}, which gives strong evidence to reject H_0 and conclude that there are more female juvenile references than male juvenile references, based on this sample. If $p_1 = p_2$ were true, then there would be almost no chance of obtaining a $\hat{p}_1$ that is so much higher than $\hat{p}_2$ with samples of these sizes.

The 90% confidence interval states that $0.275 < p_1 - p_2 < 0.537$. That is, among all such texts, the proportion of juvenile female references is from 27.5 percentage points higher to 53.7 percentage points higher than the proportion of juvenile male references.

CHAPTER

9



Inference for Two-Way Tables

9.1	Data Analysis for Two-Way Tables
9.2	Inference for Two-Way Tables
9.3	Formulas and Models for Two-Way Tables
9.4	Goodness of Fit

Introduction

In this chapter, we describe how to perform a chi-square test on data from a two-way table. We shall be testing whether there is any association between the row variable traits and the column variable traits, or whether these row and column traits are independent. We will also perform a test to decide whether data agree with a proposed distribution.

9.1 Data Analysis for Two-Way Tables

Example 9.1 Business Nonresponse. Questionnaires were mailed to 300 randomly selected businesses in each of three categorical sizes. The following data show the number of responses.

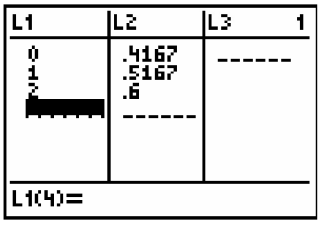
Size of company	Response	No response	Total
Small	175	125	300
Medium	145	155	300
Large	120	180	300

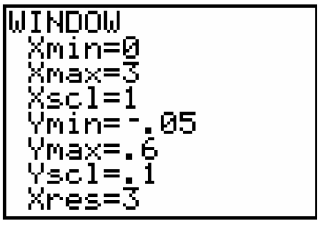
(a) What was the overall percent of nonresponse?
(b) For each size of company, compute the nonresponse percentage.
(c) Draw a bar graph of the nonresponse percents.
(d) Using the total number of responses as a base, compute the percentage of responses that came from each size of company.

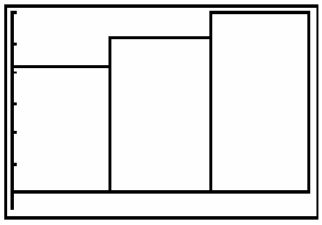
Solution. (a) To answer this question, we divide the total number of responses (125 + 155 + 180 = 460) by the total number of sent questionnaires (900). 460/900 = .5111. Overall, 51.1% of the businesses did not respond.

(b) The nonresponse rate for small companies is 125/300 = .4167 = 41.67%. For medium-sized companies, 155/300 = .5167 = 51.67% nonresponse. For large companies, 180/300 = .6 = 60% did not respond. It appears that as companies get larger, they are more likely not to respond to questionnaires.

(c) To make a (rough) bar graph of the nonresponse percents, we first enter the values **0**, **1**, and **2** into list L1 in order to represent the three types of companies, and then enter the nonresponse proportions calculated in part (b) into list L2. Next, we adjust the WINDOW (to ensure that the bars have width 1) and STAT PLOT settings for a histogram of L1 with frequencies L2, then GRAPH. Remember that proper bar graphs should not have bars connected if you try to copy this down!







(d) There were 175 + 145 + 120 = 440 total responses. Of these, 175/440 = 39.77% came from small businesses, 145/440 = 32.95% from medium-sized businesses, and 120/440 = 27.27% came from large businesses. We again see that as business size increases, the businesses are more likely not to respond, since the response rates decline.

9.2 Inference for Two-Way Tables

We continue here with an example that shows how to compute the χ^2 test of no association between the row variable and column variable in a two-way table, and obtain the matrix of expected cell counts.

TI Calculators have a built-in χ^2–Test feature (on the STAT TESTS menu) that will compute the expected counts of a random sample under the assumption that the conditional distributions are the same for each category type. To use this feature, we first must enter data from a two-way table into a matrix in the MATRX EDIT screen.

Example 9.2 Franchising Success. The following table shows the two-way relationship between whether a franchise succeeds and whether it has exclusive territory rights for a number of businesses.

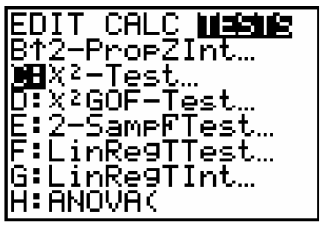
	Observed number of firms		
	Exclusive territory		
Success	Yes	No	Total
Yes	108	15	123
No	34	13	47
Total	142	28	170

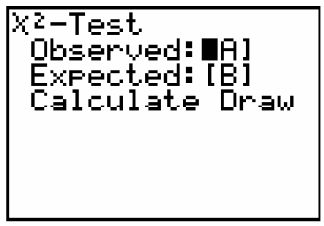
Under the assumption that there is no relationship between success and exclusive territory rights, find the expected number of successful franchises for each type of firm. And determine whether or not there is an association between franchise success and whether or not they have an exclusive territory.

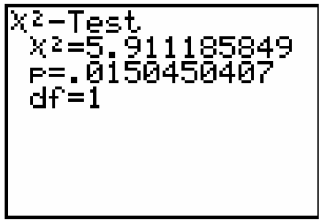
TI-83/84 Solution. First, enter the 2×2 table of data (excluding totals) into matrix [A] in the MATRIX EDIT screen. To get to this screen, press 2nd x^{-1} for the MATRIX menu, then arrow to EDIT. Press ENTER to select matrix [A]. Enter the size of the body of the matrix (rows × columns) pressing ENTER to advance the cursor. Enter the observed counts across the rows, pressing ENTER each time to advance the cursor. This menu is fussy so you must press 2nd MODE = Quit to exit the editor.

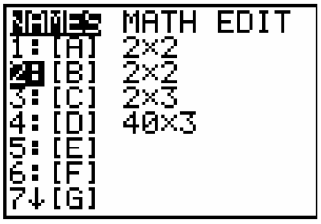
Next, bring up the χ^2–Test screen from the STAT TESTS menu, and adjust the Observed and Expected names if needed. The calculator defaults for this test are matrix [A] for Observed, and matrix [B] for Expected. If you have used another matrix, press 2nd x^{-1} for the MATRIX NAMES screen and locate the matrix name you used; press ENTER to select it. Next, press ENTER on Calculate.

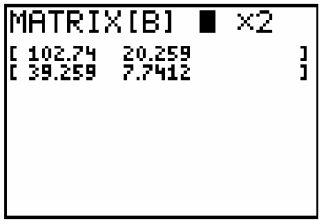






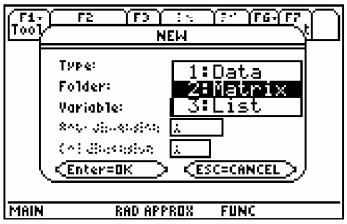


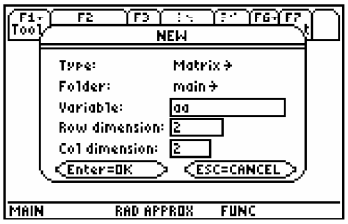


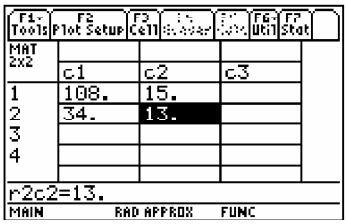


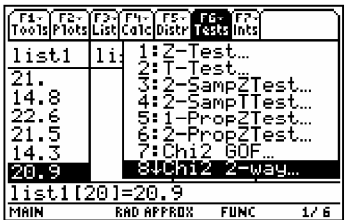
TI-89 Solution. On TI-89 calculators, the Matrix Editor is an application. Press APPS. Either press 6 or move the arrow down and press the right arrow to expand the options for what to edit. If you want to reuse a matrix name with a matrix of the same size, select option `2:Open` or select `3:New` here. You next expand the `Type:` box to select `2:Matrix`. In a similar way, select a folder (usually `Main`) and give the matrix a name (here `aa`) and its dimensions (ours is a 2×2). The matrix editor screen appears. Enter the observed counts across the rows, pressing ENTER each time to advance the cursor. This menu is fussy so you must press 2nd ESC = `Quit` to exit the editor. Return to the `Stats/List Editor` application and press 2nd F1 = F6 for the `Tests` menu. The test we want is option `8:Chi2 2-way`. Select it and use 2nd − = VAR-LINK to locate and enter your matrix name in the `Observed Mat` box. One can normally use the defaults for the `Expected` and `CompMat` matrices. As with the TI-83/84 series, there is a `Calculate` or `Draw` option for the output. Output is similar to the TI-83/84, but the first portion of the expected counts and components of the χ^2 statistic are shown. To observe these matrices fully, return to the Matrix Editor application and use the `Open` option to view them. Below, the contents of `Expmat` (the expected counts) and `Compmat` (contributions from each cell to the χ^2 statistic) are shown.

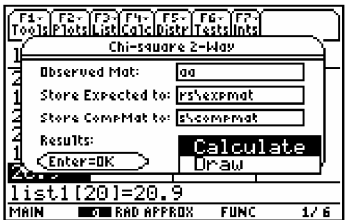


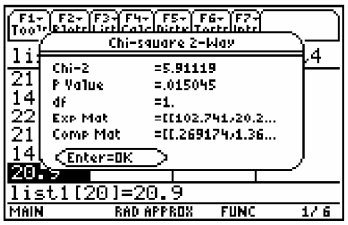


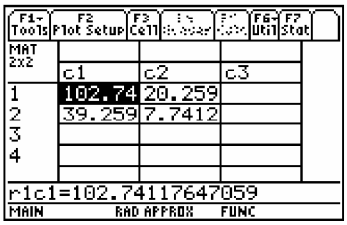


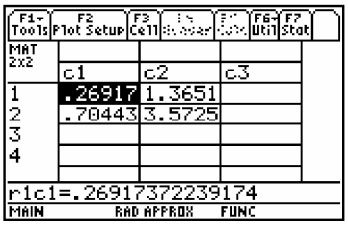










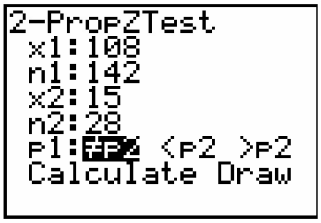


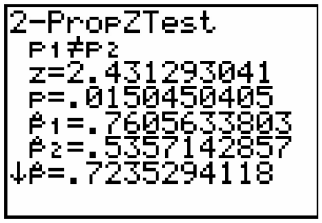
Conclusions. The χ^2–Test feature on all these calculators displays the chi-square test statistic, the *p*-value, and the degrees of freedom for the chi-square test of no association between the row and column variables. In this case, the low *p*-value of 0.015 gives strong evidence to reject the claim that there is no relationship between exclusive territory rights and franchise success. The TI-89 Compmat shows us that the largest components to the overall χ^2 statistic are from the franchises without exclusive territory. Those are most unlike what is expected under an assumption of no relationship.

If there were no relationship between success and exclusive territory rights, then we would expect 102.74 successful exclusive-territory franchises and 20.259 successful non-exclusive-territory franchises, as shown in the first row of matrix [B]. These values differ slightly from the observed values of 108 and 15 in the original data. The actual number of successful franchises in exclusive territory was somewhat higher than expected, and the actual number of successful franchises in nonexclusive territory somewhat lower than expected. It appears that having exclusive territory does help a franchise to be successful.

Comparison with the 2-PropZTest

The χ2 test for a 2×2 table is equivalent to the two-sided *z* test for H_0: $p_1 = p_2$ versus H_A: $p_1 \neq p_2$. In Example 9.2, we could let p_1 be the true proportion of all successful exclusive-territory franchises and let p_2 be the true proportion of all successful non-exclusive-territory franchises. Then $\hat{p}_1$ = 108/142 and $\hat{p}_2$ = 15/28. If we enter these values into the 2-PropZTest screen and use the alternative ≠p2, then we obtain the same *p*-value of 0.015.





9.3 Formulas and Models for Two-Way Tables

In this section, we further explain the test of no association among the row traits and the column traits versus the alternative that there is some relation between these traits.

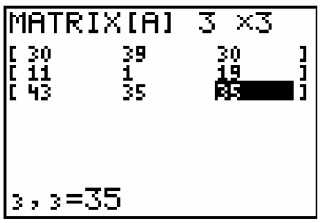
Example 9.3 Wine and Music. The two-way table that follows shows the types of wine purchased in a Northern Ireland supermarket while a certain type of music was being

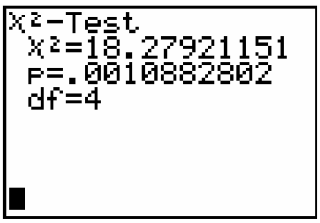
played. Test whether there is an association between the type of wine purchased and the type of music being played.

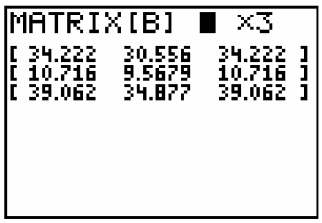
		Music		
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

Solution. Looking at the data, it appears that music may affect the type of wine purchased. Notice that only one individual bought Italian wine while French music was playing. We first enter the data (excluding totals) into a 3×3 matrix [A] in the MATRIX EDIT screen.

To see if there really is an association, we shall test the null hypothesis H_0 that there is no relation between the music and the type of wine purchased. The alternative is that there is a relation, or that the type of wine purchased does depend on the music being played. After calculating, we obtain a *p*-value of 0.001 from a test statistic of 18.2792. If there were no association between the type of wine purchased and the type of music being played, then there would be only a 0.001 probability of obtaining observed cell counts that differ so much from the expected cell counts shown in matrix [B]. Due to this low *p*-value, we can reject the null hypothesis and say that the type of music *does* affect the type of wine purchased.







Looking further at the differences between the observed and expected counts, it appears that people are more willing to buy French wine when French music is playing (the observed count was 39 and the expected count 30.556), less likely to buy Italian wine when French music is playing (our observed count was 1 and we expect 9.568 if there were no relationship), and more willing to buy Italian wine when Italian music is playing (we observed 19 and expect 10.716).

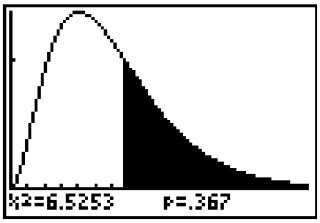
Example 9.4 Student Loans. A recent study of 865 college students found that 42.5% had student loans. The following table classifies the students by field of study and whether or not they have a loan.

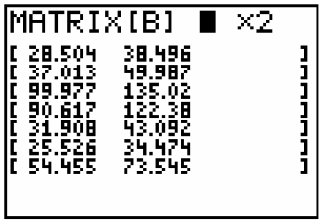
	Student loan	
Field of study	**Yes**	**No**
Agriculture	32	35
Child development and family studies	37	50
Engineering	98	137
Liberal arts and education	89	124
Management	24	51
Science	31	29
Technology	57	71

Carry out an analysis to see if there is a relationship between field of study and having a student loan.

Solution. We will test to see if the proportion of students having a student loan is the same regardless of field of study (i.e., if having a loan is independent of field). First, we enter the data into a 7×2 matrix [A] in the MATRX EDIT screen. Next, we bring up the χ^2-Test screen and designate [A] for Observed. This time selecting the Draw option, we obtain a *p*-value of 0.367 from a test statistic of 6.5253.







If having a loan were independent of field of study (i.e., if all fields had the same proportion of students with loans), then there would be a 0.367 probability of obtaining observed cell counts that differ as much from the expected cell counts. Notice that most of the expected counts in matrix [B] are close to the observed counts. Because of the high *p*-value, we can say that the observed differences are due to random chance and are not statistically significant. Thus, we will not reject the hypothesis that having a loan is independent of field of study. To further examine the data, we could compute the proportion of students in each major who have loans. For Agriculture majors, this is 32/67 = 47.8%; for Child development the proportion is 37/87 = 42.5%; for Engineering 98/235 = 41.7%; for Liberal arts 89/213 = 41.8%; Management 24/75 = 32%; Science 31/60 = 51.7%; and the proportion of Technology majors with a loan is 57/128 = 44.5%. All these proportions, with perhaps the exception of Management and Science, are very similar.

9.4 Goodness of Fit

TI-84 and -89 calculators have a built-in function to perform the goodness of fit for a discrete distribution. The commands for a TI-83 are simple enough, but we provide a small program below that also will accomplish the task.

Before executing the **FITTEST** program that follows, enter the specified proportions into list L1 and enter the observed cell counts into list L2. The expected cell counts are computed and stored in list L3, and the individual contributions to the chi-square test statistic are stored in list L4. The program displays the test statistic and the *p*-value.

The FITTEST Program

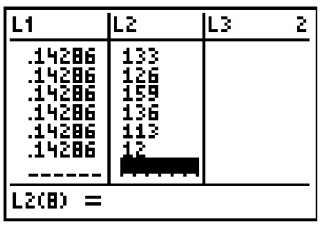
Program:FITTEST	:1-χ2cdf(0,X,dim(L2)-1)STO▶P
:sum(L2)STO▶N	:ClrHome
:N*L1STO▶L3	:Disp "CHI SQ STAT"
:(L2-L3)^2/L3STO▶L4	:Disp X
:sum(L4)STO▶X	:Disp "P VALUE"
	:Disp P

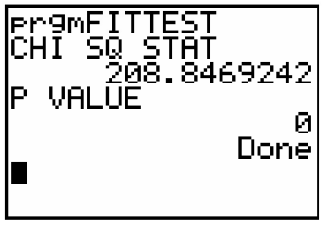
Example 9.5 Cell Phones and Accidents. The following table gives the number of motor vehicle collisions by drivers using a cell phone broken down by days of the week over a 14-month period. Are such accidents equally likely to occur on any day of the week? It appears from the table that weekends have fewer accidents of this type; is this real or just randomness at work?

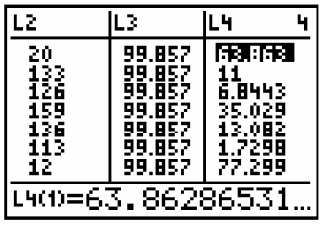
Number of collisions by day of the week

Sun.	Mon.	Tue.	Wed.	Thurs.	Fri.	Sat.	Total
20	133	126	159	136	113	12	699

TI-83 Solution. If each day were equally likely, then 1/7 of all accidents should occur on each day. To test the fit of this distribution, we shall use the **FITTEST** program. We first enter **1/7** seven times into list L1 (the calculator displays a decimal equivalent) to specify the expected distribution, and enter the given frequencies from the table into list L2. Next, we run the **FITTEST** program to obtain a *p*-value of 0 from a chi-square test statistic of 208.8469.

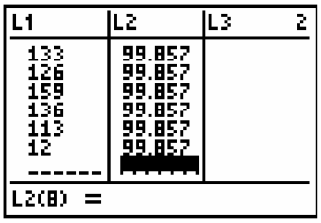


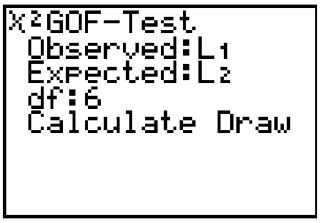


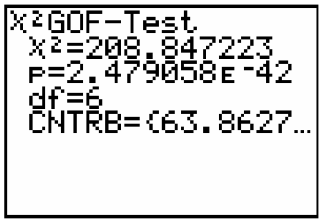


TI-84/89 Solution. Both of these calculators have built-in chi-square goodness-of-fit tests and they function similarly. In both cases, we enter the observed and expected counts into two lists. The expected count under the null hypothesis (accidents are equally likely

each day of the week) is 1/7*699 = 99.857. Our observed counts have been entered into L1 and the expected counts in L2. From the STAT TESTS menu select option χ2GOF-Test (Chi2 GOF on a TI-89). Enter the list names for your observed and expected counts and the degrees of freedom for the test which are $k - 1$, where k is the number of categories. Since there are 7 days per week, for this example $k = 7$, so we have df = 6. The TI-84 output seen below shows the contributions to the χ^2 statistic. To see all of them, use the right-arrow key. The TI-89 creates a new list of components which can be viewed in the List editor.



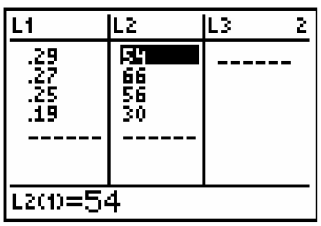


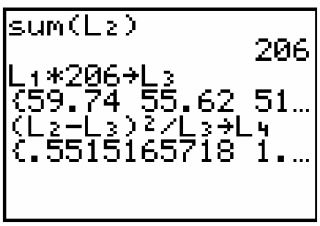


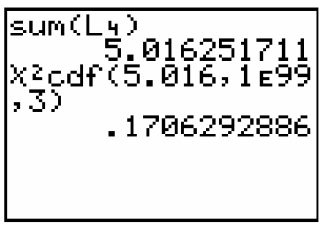
Conclusions. If these accidents were equally likely to occur on any day of the week, then there would be no chance of obtaining a sample distribution that differs so much from the expected counts of 1/7*699 = 99.857 for each day. So we can reject the claim that accidents are equally likely on each day. Looking at the components, we clearly see that weekends are much lower than expected; Wednesday is also higher than expected.

Example 9.6 Sampling College Students. At a particular college, 29% of undergraduates are in their first year, 27% in their second, 25% in their third, and 19% are in their fourth year. But a random survey found that there were 54, 66, 56, and 30 students in the first, second, third, and fourth years, respectively. Use a goodness-of-fit test to examine how well the sample reflects the college's population.

Solution. Using commands on a TI-83,we shall to test the goodness of fit to examine how well the sample reflects the college's population. We have entered the stated proportions **0.29**, **0.27**, **0.25**, and **0.19** into list L1 and the observed frequencies into list L2. On the home screen, we first sum L2 to obtain the total number in the sample, then multiply the total by the proportions in L1 to obtain the expected counts which are stored in L3. The components of the χ^2 statistic, $(O - E)^2/E$ are stored in L4. We sum L4 to obtain the statistic and use χ2pdf from the DISTR menu to find the p-value. We obtain a p-value of 0.17061 from a chi-square test statistic of 5.016. For the given distribution, there is about a 17% chance of obtaining observed sample counts that differ as much as these do from the expected counts. We conclude that the sample does not differ significantly from what was expected.

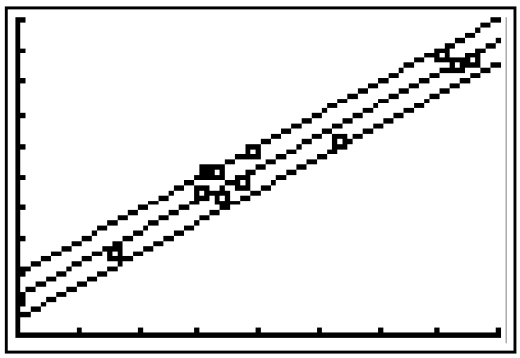






CHAPTER

10



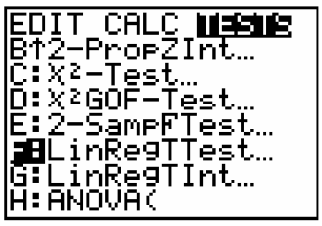
Inference for Regression

10.1	Simple Linear Regression
10.2	More Detail about Simple Linear Regression

Introduction

In this chapter, we provide details on using TI calculators to perform the many difficult calculations for linear regression. In particular, we again find and graph the least-squares regression line and compute the correlation. We can then perform a t test to check the hypothesis that the correlation (or, equivalently, the regression slope) is equal to 0. We will also show how to compute a test statistic and find a p-value when the hypothesized slope is something other than 1. We also provide a program that computes confidence intervals for the regression slope and also computes a prediction interval for a future observation and a confidence interval for a mean response.

10.1 Simple Linear Regression



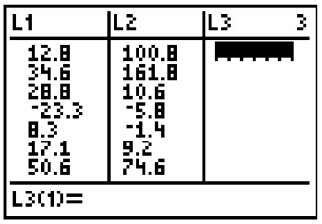
We begin by demonstrating the LinRegTTest feature (item E) from the STAT TESTS menu that will compute a least-squares regression line while simultaneously testing the null hypothesis that the regression slope equals 0. Notice that my TI-84 calculator also has a LinRegTInt feature. This is also available on a TI-89. We'll discuss that (and a TI-83 program to perform this) later.

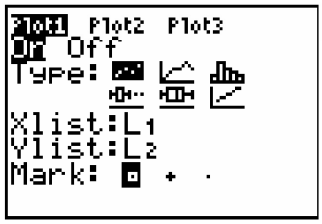
Example 10.1 Stock and Bond Investing. Here are data on the net new money (in billions of dollars) flowing into stock and bond mutual funds from 1985 to 2000. Negative values indicate more money went *out* of the market than went in. We'd like to find out whether or not the relationship is statistically significant. In other words, are the two related, and if so, how?

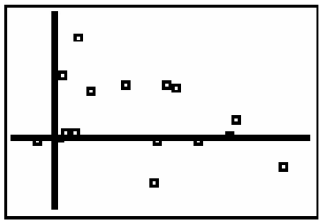
Year	1985	1986	1987	1988	1989	1990	1991	1992
Stocks	12.8	34.6	28.8	–23.3	8.3	17.1	50.6	97.0
Bonds	100.8	161.8	10.6	–5.8	–1.4	9.2	74.6	87.1

Year	1993	1994	1995	1996	1997	1998	1999	2000
Stocks	151.3	133.6	140.1	238.2	243.5	165.9	194.3	309.0
Bonds	84.6	–72.0	–6.8	3.3	30.0	79.2	–6.2	–48.0

Solution. We begin by making a scatterplot as explained in Section 2.1 by entering the data into lists and defining a scatterplot on the STAT PLOT screen that uses stocks as the predictor (X) variable and Bonds as the response (Y) variable. Press ZOOM 9 to display the plot. The relationship is fairly linear, but not extremely strong, since there is a good bit of scatter in the plot. The relationship is a decreasing one; if more money went to stocks, less went to bonds.







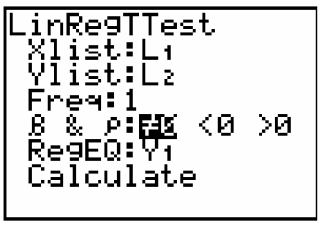
In Chapter 2, we showed how to compute and graph the least-squares line using the LinReg(a+bx) command from the STAT CALC menu. But now we shall use LinRegTTest feature from STAT TESTS menu. As before, we first should make sure that the DiagnosticOn command has been entered from the CATALOG.

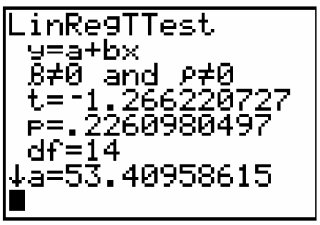
With LinRegTTest, we can find the least-squares line while at the same time testing the null hypothesis that the slope β_1 is equal to 0. We will use the alternative hypothesis

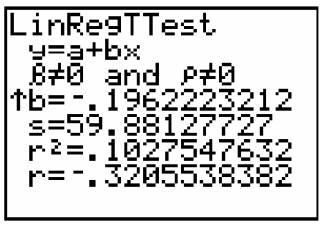
that $\beta_1 \neq 0$. This test is equivalent to testing the null hypothesis that the population correlation ρ is equal to 0 with an alternative that $\rho \neq 0$.

After bringing up the LinRegTTest screen, adjust the settings for L1 versus L2 (or to whatever lists contain the data) and enter the alternative ≠Ø. To enter Y1 for RegEQ, scroll down to the right of RegEQ, press [VARS], scroll right to Y-VARS, press [ENTER] for Function and then press [ENTER] for Y1. Then press [ENTER] on Calculate.

We obtain the regression line y = a + bx (or $y = \hat{\beta}_0 + \hat{\beta}_1 x$), which is also stored into Y1. Here the equation rounds to y = 53.4096 – 0.1962x. Press [GRAPH] to see the plot with the regression line.







The p-value for the t test is given as 0.226 from a t statistic of –1.26622. If β_1 were equal to 0, then there would be a 22.6% chance of obtaining a value for b as low as –0.1962, or of obtaining a correlation as low as r = –0.32, with a sample of this size. This rather high p-value means that we do *not* have statistically significant evidence that there is some straight-line relationship between the flows of cash into bond funds and stock funds. In other words, we do not have enough evidence to reject that $\beta_1 = 0$ or to reject that $\rho = 0$.

Hypotheses Other Than 0

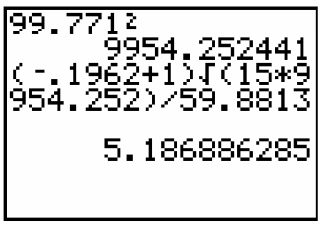
In certain circumstances, one might be interested in a null hypothesis different from $H_0 : \beta = 0$. The general t statistic for a test of this type is

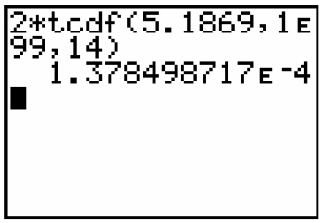
$$t = \frac{b_1 - \beta_{10}}{s(b_1)} = \frac{b_1 - \beta_{10}}{s_e \sqrt{\dfrac{1}{\sum (x - \bar{x})^2}}} = \frac{(b_1 - \beta_{10})\sqrt{(n-1)s_x^2}}{s_e}$$

with n – 2 degrees of freedom. In this formula, β_{10} is the hypothesized slope, s_e is the standard deviation of the residuals around the regression line, and s_x^2 is the sample variance of the predictor (x) variable. This t statistic has n – 2 degrees of freedom.

Example 10.2 Another Hypothesis About the Slope. In Example 10.1, we might be concerned that in any given month there are only so many dollars available to invest. Let's call that D. If investors were to split their money between stocks and bonds, we

would have an equation $S + B = D$. This would argue that the regression relationship should be of the form $B = D - S$, and we should have slope $\beta_1 = -1$. We will test this as H_0 against H_A: $\beta_1 \neq -1$. We had $b_1 = -0.1962$, $s_e = 59.8813$ (given as s on the calculator output), and using `1-Var Stats`, we find $s_x^2 = 99.771^2 = 9954.252$. Putting this together with the fact that we had $n = 16$ data points, we find a t statistic of 5.1869. To find the p-value for this two-sided test, we will double the area above $t = 5.1869$ under the t distribution curve with 14 degrees of freedom. Our p-value is 0.00014, so we have sufficient evidence based on this sample that there is not an inverse relationship between net investments in stocks and bonds (the true slope is not –1.)





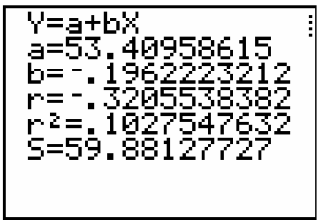
Confidence Intervals in Regression Inference

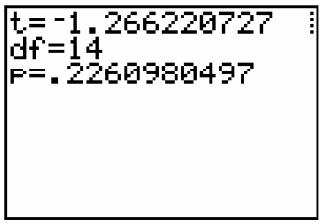
We provide a program called **REGINF** (the listing is at the end of this chapter) that will compute confidence intervals for the slope and for the intercept of the linear regression model $y = \beta_0 + \beta_1 x$, as well as confidence intervals for the mean value of y at a given x and prediction intervals y given a new value of x. TI-84 calculators have a built-in function to compute confidence intervals for the slope (it is shown on the first line of the output, the rest is similar to `LinRegTTest`); TI-89 calculators will find all these intervals (but only at 95% confidence, since that is the most common value) given the right input. The program below can be used on TI-83/84 calculators for more general confidence and prediction intervals.

Example 10.3 Stocks and Bonds Continued. Using the data from Example 10.1, find the 90% confidence intervals for the slope β_1, and intercept β_0 of the linear regression model.

TI-83/84 Program Solution. With the data already entered into lists `L1` and `L2`, bring up the **REGINF** program. The program first asks for the list names containing your data. It then displays the regression output we have seen before from `LinRegTTest`. Notice the scrolling dots at the upper right of the screen. Press [ENTER] to continue on.

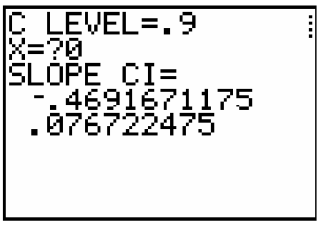


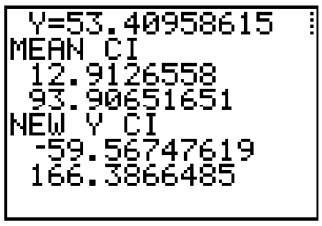




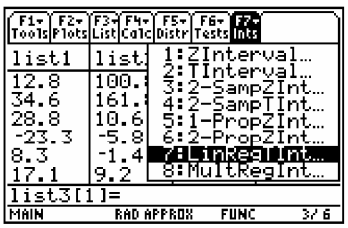
You are now asked for the confidence level. Here we want 90%, so enter .9. You are now asked what x value we are interested in. Since we are interested in the intercept, enter 0. Press [ENTER] to continue. We first find a 90% confidence interval for the slope. Based on these data we are 90% confident the true slope for predicting net bond dollars based on net stock dollars is between –0.469 and 0.077. Notice that our original null hypothesis H_0: $\beta_1 = 0$ is contained in the interval. Press [ENTER] again and we will find the value predicted from the line at our desired x value (notice it is exactly the intercept here) and a confidence interval for the mean response y at that x and a prediction interval for a new observation at x. Press [ENTER] again and the program will display a "Done" message.

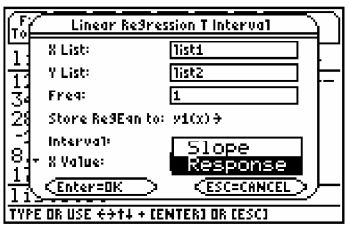


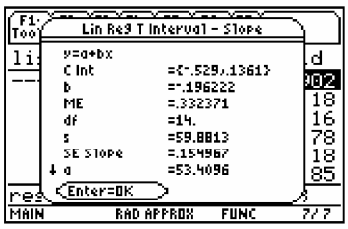




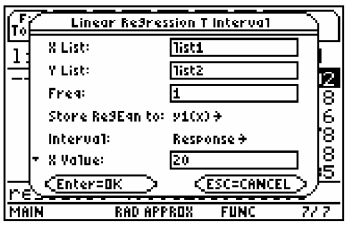
TI-89 Solution. The data have been entered in `list1` and `list2`. From the [2nd][F2]=[F7] `Ints` menu, select option `7:LinRegTInt`. Use [2nd][-] = [VAR-LINK] to enter the list names. Store the regression equation (if desired). We can use the right arrow to expand the `Interval` box to select the type of interval: `Slope` or `Response`. If we select for a Slope, press [ENTER] to perform the calculations; if for a Response, you can enter the particular x value of interest (for the intercept this would be 0). Here we see that a 95% confidence interval for the slope is (–0.529, 0.126). We also see the standard error of the slope, $s(b_1)$, labeled as `SE Slope`. Press the down arrow to find r and r^2.

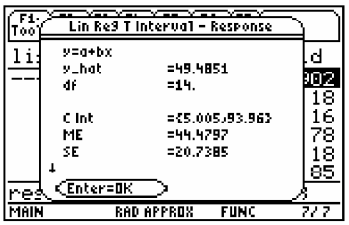


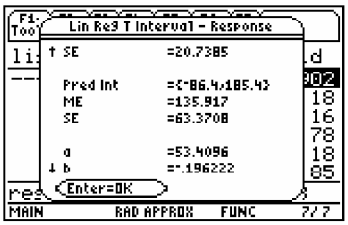




Suppose we wanted to find confidence and prediction intervals for the amount of money going into bonds when 20 billion dollars is going to stocks. Use the same as above, but select `Response` as the `Interval` type and type 20 in the `X Value` box. The first portion of the output gives the $\hat{y}$ prediction from the line. Further down, we see that, with 95% confidence, on average there should be between 5.005 and 93.96 billion dollars going into bonds when 20 billion dollars is going into stocks. We are also explicitly told the margin of error and the standard error of the confidence interval. Press the down arrow to find the prediction interval information. For a particular month with 20 billion dollars going into stocks, with 95% confidence, somewhere between –86.4 and 185.4 billion dollars will go into bonds.





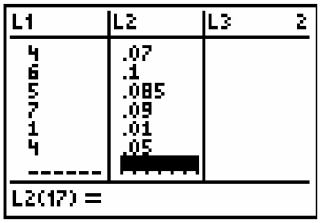


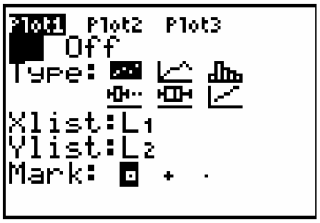
Example 10.4 Beer and Blood Alcohol. Several years ago a study was conducted at The Ohio State University in which 16 student volunteers were randomly assigned to drink a number of cans of beer. The students were equally divided between men and women and varied in weight and normal drinking habits. Thirty minutes after they finished drinking, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood. The data are displayed below. Investigate the relationship between number of beers drunk and blood alcohol level.

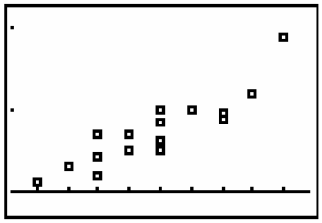
Student	1	2	3	4	5	6	7	8
Beers	5	2	9	8	3	7	3	5
BAC	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06

Student	9	10	11	12	13	14	15	16
Beers	3	5	4	6	5	7	1	4
BAC	0.02	0.05	0.07	0.10	0.085	0.09	0.01	0.05

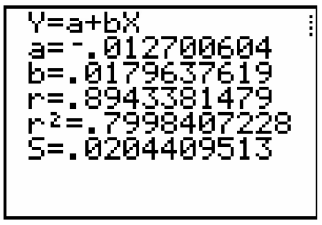
Solution. First, plot the data. Is the relationship linear enough for our regression to make sense? The data have been entered into lists L1 and L2. We define the scatterplot and press ZOOM 9 to display it. The plot is very linear, but the individual represented at the upper right (9 beers, BAC 0.19) might be influential to this regression.

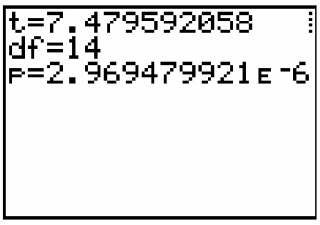


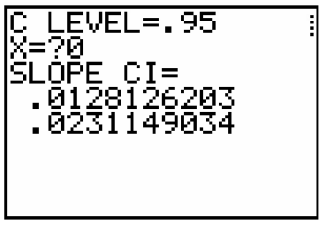




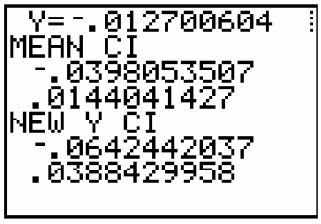
Using program **REGINF**, we find the following: The regression equation is BAC = –0.127 + 0.180*Beers. About (r^2) 80% of the overall variation in BAC is accounted for by the number of beers consumed. The standard deviation of the residuals around the line is s = 0.0204. The relationship is highly significant. With a p-value of 0.00000297, we're convinced based on this sample that there is a linear relationship between the number of beers consumed and BAC. Further, we are 95% confident the mean blood alcohol level increases between 0.0128 and 0.0231 per beer.







Last, the intercept is *not* significantly different from 0, since 0 is included in the confidence interval for the value when $x = 0$. This makes sense; when a person hasn't had anything alcoholic to drink, they should have a BAC of 0.

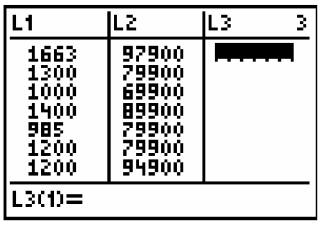


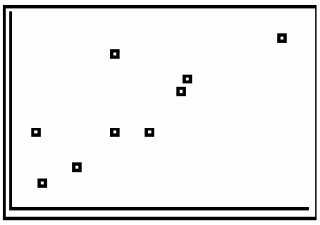
Example 10.5 Augusta House Prices. Not too long ago, the author of this manual was involved in the purchase of a house for her daughter in Augusta, GA. The sizes of the houses in the first group we looked at (all older homes) and their asking prices are:

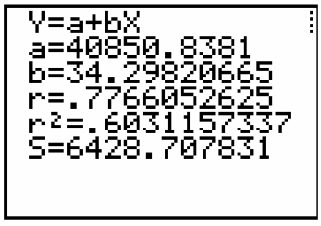
Sq. Ft.	1663	1300	1000	1400	985	1200	1200	1100	1384
Price ($)	97,900	79,900	69,900	89,900	79,900	79,900	94,900	73,000	87,595

Find a linear regression model to relate the size of the house to the asking price, and a confidence interval for the slope. Find an interval estimate for the average price of a house with 1200 square feet of space.

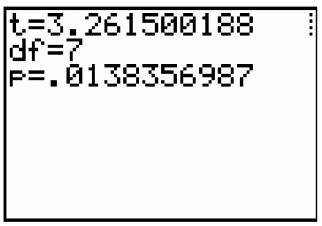
Solution: Enter the data into two lists, plot the data, then use program **REGINF** to find the regression equation and the confidence interval for the slope. First, we notice that there appears to be a strong relationship between the size of the house and the asking price; however, there also seem to be two outliers which may affect our results.

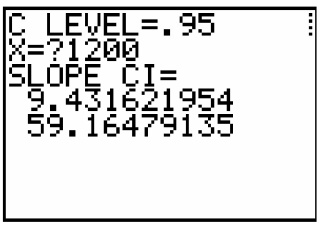


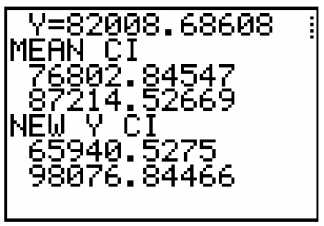




The regression is $Asking\$ = 40850.84 + 34.30 * sqft$. Further, we see that the size of the house explains about 60% of the variation in asking price (other lurking variables like age, materials, location, will explain more).







From confidence interval for the slope, we see that we are 95% confident the average cost per square foot in Augusta at that time (for houses of this type) was somewhere between \$9.43 and \$59.16. Based on this sample, the average asking price of older homes in Augusta at this time with 1200 square feet should have been between \$76,802.85 and \$87,214.53 with 95% confidence.

Intervals Without a Program

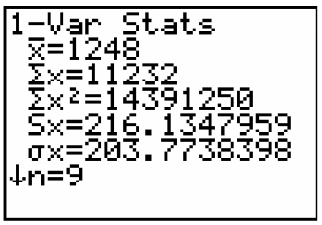
Keying in a program can be tedious, if you do not have access to a TI-connect cable. Here, we demonstrate how to use quantities available with standard calculator output to create confidence or prediction intervals of your own. We will use two basic formulas:

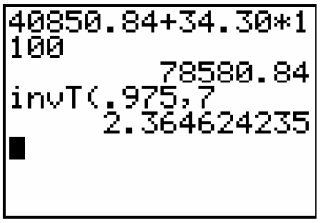
Confidence interval for the mean response at x: $\hat{y} \pm t^* s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x-\bar{x})^2}}$

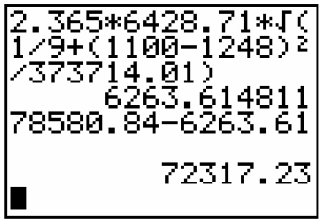
Prediction interval for a new response at x: $\hat{y} \pm t^* s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x-\bar{x})^2}}$

Example 10.6. More on House Prices. Based on the model found in Example 10.5, what should be the average price for an older home in Augusta with 1100 square feet? Use a 95% confidence interval estimate.

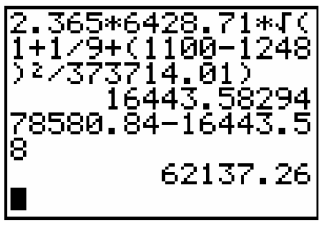
Solution. As demonstrated in Example 10.2, we can use `1-Var Stats` on the list containing our x variable to find both `Sx` and $\bar{x}$. Since $\sum (x-\bar{x})^2 = (n-1)s_x^2$ we have $\sum (x-\bar{x})^2 = 8*216.1348^2 = 373714.01$. We also see n = 9. From calculating the regression, we have $s_e = 6428.71$. Using `invT` from the `DISTR` menu (or a table), we find t^* = 2.365. Also, using our equation for the regression, we have a point estimate of $\hat{y} = 40850.84 + 34.20*1100 = \$78,580.84$. Putting it all together, we find a margin of error for this confidence interval is 6263.61. The lower end of the confidence interval is \$72,317.23, and we similarly find the upper end is \$84,844.45.







Example 10.7 Still More House Prices. Suppose the real estate agent told us about a "really nice" home with 1100 square feet of space. What should its asking price be, based on this model with 95% confidence?



Solution. One particular home is not an average. In this case, we need a prediction interval for a new observation. We found all the needed pieces in Example 10.5. All we need to do is add another 1 under the square root to account for the fact that individuals are more variable than means. The new margin of error is \$16,443.58. The lower end of the 95% prediction interval is \$62,137.26. Changing the minus sign to a plus, we find the upper end of the prediction interval is \$95,024.42.

10.2 More Detail about Simple Linear Regression

Analysis of variance (ANOVA) is another method to test the null hypothesis H_0: $\beta_1 = 0$, with an alternative H_A: $\beta_1 \neq 0$. Although really more useful in a true ANOVA or multiple regression setting, in the case of simple linear regression it has a use and relation to a *t* test similar to that of using a χ^2 test for two proportions. The **REGANV** program on page 118 performs such a test, stores an ANOVA table into lists L4, L5, and L6, and displays the associated *F*-statistic and *p*-value. Before executing the **REGANV** program, we must enter paired data into lists L1 (*x*) and L2 (*y*).

Note: The ANOVA table generated by the **REGANV** program is not the same one that will be generated by the calculator's built-in ANOVA(command that is used for testing whether or not several populations have identical means. That command will be explained in Chapter 12.

The REGANV Program for Linear Regression ANOVA

```
PROGRAM:REG3
:LinReg(a+bx) L1,L2
:sum(seq((a+bL1(I)-ȳ)²,
 I,1,dim(L1)))→A
:sum(seq((RESID(I))²,I,1,
 dim(L1)))→B
:sum(seq((L2(I)- ȳ)²,
 I,1,dim(L1)))→C
:(n-2)A/B→F
:1-Fcdf(0,F,1,n-2) →P
:ClrList L4,L5,L6
:1→L4(1):A→L5(1):A→L
6(1)
:n-2→L4(2):B→L5(2):B/(n-2)
→L6(2)
:n-1→→L4(3):C→L5(3):C/(n-1)
→L6(3)
:ClrHome
:Disp "F-STAT, P-VAL"
:Output(2,2,{round(F,3),round(P,4)})
:Output(4,1,"SEE L4, L5, L6")
:Output(5,5,"DF, SS, MS")
:Output(6,3,"M")
:Output(7,3,"E")
:Output(8,3,"T")
:Stop
```

Example 10.8 More Stocks and Bonds. Consider the data from Example 10.1 on the net new money (in billions of dollars) flowing into stock and bond mutual funds from 1985 to 2000.

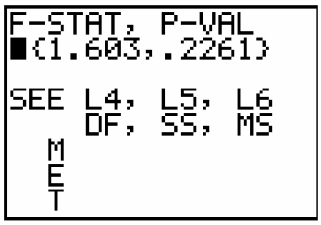
Year	1985	1986	1987	1988	1989	1990	1991	1992
Stocks	12.8	34.6	28.8	–23.3	8.3	17.1	50.6	97.0
Bonds	100.8	161.8	10.6	–5.8	–1.4	9.2	74.6	87.1

Year	1993	1994	1995	1996	1997	1998	1999	2000
Stocks	151.3	133.6	140.1	238.2	243.5	165.9	194.3	309.0
Bonds	84.6	–72.0	–6.8	3.3	30.0	79.2	–6.2	–48.0

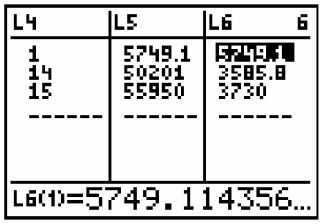
(a) Construct the ANOVA table.
(b) State and test the hypotheses using the ANOVA F-statistic.
(c) Give the degrees of freedom for the F-statistic for the test of H_0.
(d) Verify that the square of the t statistic for the equivalent t test is equal to the F-statistic in the ANOVA table.

Solution. (a) To construct an ANOVA table with the **REGANV** program, we must first enter our data into lists L1 (for stocks) and L2 (for bonds). After doing so, bring up and execute the program. (There is nothing to input, so just press ENTER.)

The program displays the F-statistic and p-value. The ANOVA table is stored in lists L4, L5, and L6 and contains the degrees of freedom (DF), the sum of squares (SS), and the mean square (MS) for each of the model (M), error (E), and total (T). Notice that the p-value for the F test is exactly what we found in Example 10.1 with LinRegTTest (page 111).



The ANOVA table is displayed at right. The model has 1 degree of freedom here since we had one predictor variable. Total degrees of freedom in the last line is $n - 1$. Its sum of squares would be the numerator in computing the variance of the y variable in the regression, and Mean Square is the variance of y. If we divide SSM by SST we have 5749.1/55950 = 0.10275 which was the r^2 found in Example 10.1.



(b) The ANOVA test is about the linear regression slope β_1. We test the null hypothesis H_0: $\beta_1 = 0$ with the alternative H_A: $\beta_1 \neq 0$. With the above p-value of 0.2261, we do not have strong evidence in this case to reject H_0 in favor of the alternative.

(c) The degrees of freedom for the F-statistic are given by 1, $n - 2$, which in this case is 14. This number is the same as the degrees of freedom of the error (E) displayed in the ANOVA table.

(d) The LinRegTTest was used in Example 10.1. The t statistic was computed as

$t = -1.266220727$. If we square this value, then we obtain 1.603314929, which is the actual value of the displayed (rounded) F-statistic from the ANOVA test.

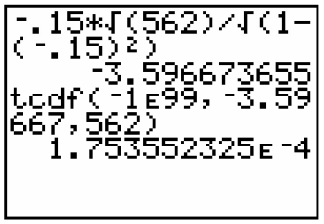
Sample Correlation and the *t* Test

One may be required to perform a correlation t test without an actual data set, but instead by using only the values of the sample correlation r and the sample size n. Although TI calculators do not have a built-in procedure for this type of test, the t statistic and p-value are easily calculated in this case. Here the test statistic, which follows a $t(n-2)$ distribution, is given by

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Example 10.9 Avoidance of New Foods. In a study of 564 children who were 2 to 6 years of age, the relationship of food neophobia (avoidance of unfamiliar foods) and the frequency of consumption gave a correlation of $r = -0.15$ for meat. Perform a significance test about the correlation of meat neophobia and the frequency of meat consumption among children 2 to 6 years of age.

Solution. We shall test H_0: $\rho = 0$ versus H_A: $\rho < 0$. The rejection region is a left tail; thus the p-value will be the left-tail probability of the $t(n-2) = t(562)$ distribution. We compute the t statistic and p-value "manually" by entering $-.15*\sqrt{562}/\sqrt{1-(-.15)^2}$ to obtain $t = -3.59667$. Next, we compute the p-value $P(t(562) \le -3.59667)$ with the `tcdf(` command from the `DISTR` menu. We enter `tcdf(E1E99,-3.59667,562)` to obtain a p-value of 1.75×10^{-4}.



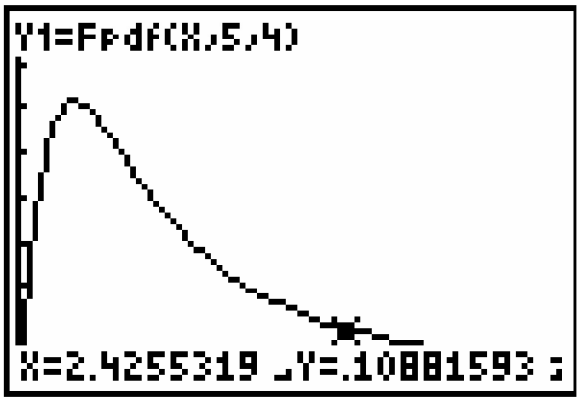
If the true correlation were 0, then there would be only a 0.000175 probability of obtaining a sample correlation as low as $r = -0.15$ with a random sample of size 564. We have statistical evidence to reject H_0 in favor of the alternative that $\rho < 0$. We note that for a two-sided alternative, then our p-value would be $2*(1.75\times10^{-4}) = 3.5\times10^{-4}$, which would still be small enough for us to reject H_0.

The REGINF Program for Linear Regression Inference

```
:Input "X LIST=",LX
:Input "Y LIST=",LY
:FnOff
:LinRegTTest LX,LY,0,Y1:
:σx²n→V:s→S:x̄→M:n→N
:ClrHome
:Output(1,2,"Y=a+bX"
:Output(2,2,"a="
:Output(2,4,a
:Output(3,2,"b="
:Output(3,4,b
:Output(4,2,"r="
:Output(4,4,r
:Output(5,2,"r²="
:Output(5,5,r²
:Output(6,2,"S="
:Output(6,4,S
:Pause
:df→K
:ClrHome
:Output(1,1,"t="
:Output(1,3,t
:Output(2,1,"df="
:Output(2,4,K
:Output(3,1,"p="
:Output(3,3,p
:Pause
:ClrHome
:Input "C LEVEL=",C
:Prompt X
:K+1→K:b→A
:TInterval 0, √(K),K,C
:upper→T
:A-T*S/√(V)→P
:A+T*S/√(V)→Q
:Output(3,1,"SLOPE CI="
:Output(4,2,P
:Output(5,2,Q
:Pause
:Y1-ST√(N⁻¹+(X-M)²/V)→P
:Y1+ST√(N⁻¹+(X-M)²/V)→Q
:Y1-ST√(1+N⁻¹+(X-M)²/V)→J
:Y1+ST√(1+N⁻¹+(X-M)²/V)→U
:ClrHome
:Output(1,2,"Y="
:Output(1,4,Y1
:Output(2,1,"MEAN CI"
:Output(3,2,P
:Output(4,2,Q
:Output(5,1,"NEW Y CI"
:Output(6,2,J
:Output(7,2,U
:Pause
:ClrHome
```

CHAPTER

11



Multiple Regression

11.1	Inference for Multiple Regression
11.2	A Case Study

Introduction

In this chapter, we demonstrate how to use a program for TI-83 and -84 calculators to calculate a multiple linear regression model. We also show how to do this with built-in functions on a TI-89.

11.1 Inference for Multiple Regression

The **MULREG** program below computes the regression coefficients and an ANOVA table for the multiple linear regression model $\mu_Y = \beta_0 + \beta_x x_1 + ... + \beta_k x_k$. The program stores the regression coefficients in L1, their standard errors in L2, individual t statistics in L3, and p-values in L4. The squared multiple correlation coefficient (r^2), the F-statistic, the p-value, and the standard deviation of the residuals are also displayed along with the entries in the ANOVA table. The program can also be used to find prediction and confidence intervals for the response given a set of predictor values.

The MULREG Program

```
Disp "DATA IN COLS","OF [A]"
dim([A])→θ1:Lθ1(1)→N:Lθ1(2)→L:L-1→K
{N,1}→dim([B]):Fill(1,[B])
augment([B],[A])→[B]
[B]^T[B]→[D]
seq([D](I,1),I,2,L+1)/N→Lθ1
{L,L}→dim([C]):Ans→dim([E])
{N,1}→dim([C])
Input "RESPONSE COL=",R
For(I,1,N):[A](I,R)→[C](I,1):End
If R≤K:Then
For(J,R+1,L):For(I,1,N)
[B](I,J+1)→[B](I,J)
End:End
[B]^T[B]→[D]
End
{L,L}→dim([D]):[D]x^-1→[D]
{N,L}→dim([B])
[B]^T[C]:[D]Ans→[E]
Matr►list([E],Lθ1)
Disp "COEF Lθ1=":Pause Lθ1
[B][E]→[B]
Matr►list([C],LY)
Matr►list([B],LYP)
mean(LY)→Y
sum((LY-Y)^2)→T
sum((LY-LYP)^2)→E
LYP→L5
LY-LYP→L6
DelVar LY:DelVar LYP
N-L→M:T-
E→R:R/K→Q:E/M→D:√(D)
→S:Q/D→F
S√(seq([D](I,I),I,1,L))→Lθ2
Disp "STDEV Lθ2=":Pause Lθ2
Lθ1/Lθ2→Lθ3
Disp "T-RATIO Lθ3=":Pause Lθ3
Disp "COEF P Lθ4="
1-2seq(tcdf(0,abs(Lθ3(I)),M),I,1,L)→Lθ4
Pause Lθ4
1-(N-1)*D/T→A
ClrHome
Output(1,1,"S="
Output(1,9,S
Output(2,1,"R^2="
Output(2,9,R/T
Output(3,1,"R^2ADJ="
Output(3,9,A
Output(4,1,"REG DF="
Output(4,9,K
Output(5,1,"ERR DF="
Output(5,9,M
Output(6,1,"TOT DF="
Output(6,9,N-1
Pause
ClrHome
Output(1,1,"SS REG="
Output(1,9,R
Output(2,1,"SS ERR="
Output(2,9,E
Output(3,1,"SS TOT="
Output(3,9,T
Output(4,1,"MS REG="
Output(4,9,Q
Output(5,1,"MS ERR="
Output(5,9,D
Output(6,1,"F=  "
Output(6,9,F
Output(7,1,"P-VAL="
1-Fcdf(0,F,K,M)→P
round(P,5)→P
Output(7,9,P
Pause
ClrHome
Lbl C
```

```
Menu("PREDICT Y?","YES",A,"QUIT",B)
Lbl A
Input "X{}=",Lθ5
Input "CONF. LEVEL=",C
{L,1}→dim([C])
1→ [C](1,1)
For(I,1,K):Lθ5(I)→ [C](I+1,1):End
[E]^T[C]
Ans(1,1)→Y
ClrHome
round(Y,5)→Y
Output(1,1,"Y HAT="
Output(1,8,Y
[C]^T[D][C]
Ans(1,1)→V
round(S√(V),5)→T
Output(2,1,"S(FIT)="
Output(2,8,T
TInterval 0,√(M+1),M+1,C
upper→T
Output(3,1,"C.I.="
Y+TS√(V)→D
Y-TS√(V)→E
Output(4,3,E
Output(5,3,D
Output(6,1,"P.I.="
Y-TS√(1+V)→E
Y+TS√(1+V)→D
```

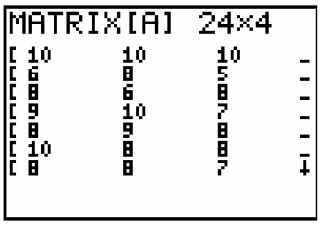
To execute the program, we first must enter our sample data into matrix [A] with the dependent (*Y*) variable in the last column.

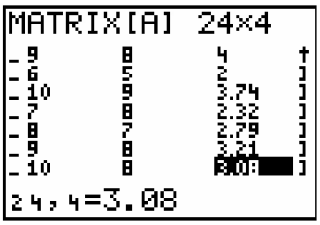
Example 11.1 Predicting Grades. Consider this data from a sample of 24 students at a large university that uses a 4.0 GPA grade scale. Run a multiple regression analysis for predicting the GPA from the three high school grade variables (Math, Science, and English where 10 = 'A', 9 = 'A–', 8='B+', etc).

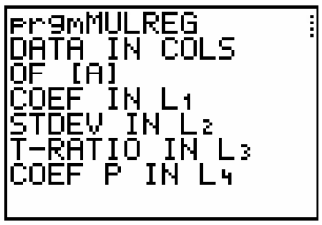
OBS	GPA	HSM	HSS	HSE	SATM
1	3.32	10	10	10	670
2	2.26	6	8	5	700
3	2.35	8	6	8	640
4	2.08	9	10	7	670
5	3.38	8	9	8	540
6	3.29	10	8	8	760
7	3.21	8	8	7	600
8	2.00	3	7	6	460
9	3.18	9	10	8	670
10	2.34	7	7	6	570
11	3.08	9	10	6	491
12	3.34	5	9	7	600
13	1.40	6	8	8	510
14	1.43	10	9	9	750
15	2.48	8	9	6	650
16	3.73	10	10	9	720
17	3.80	10	10	9	760
18	4.00	9	9	8	800
19	2.00	9	6	5	640
20	3.74	9	10	9	750
21	2.32	9	7	8	520
22	2.79	8	8	7	610
23	3.21	7	9	8	505
24	3.08	9	10	8	559

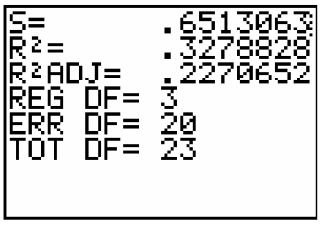
TI-83/84 Solution. We use the model $\mu_{GPA} = \beta_0 + \beta_1 x_{HSM} + \beta_2 x_{HSS} + \beta_3 x_{HSE}$, which is of the form $\mu_Y = \beta_0 + \beta_x x_1 + \beta_2 x_2 + \beta_3 x_3$. In this case, there are 24 sample points and four variables. We must first enter the data into matrix [A], much as we have been entering data into lists. Matrix [A] should have 24 rows and 4 columns. However, to execute the program correctly, the dependent variable must be in the *last* column.

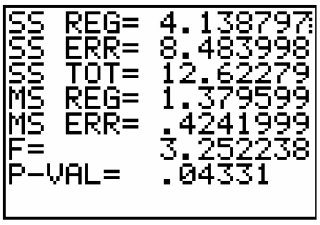
In the MATRIX EDIT screen, enter the dimensions of matrix [A] as 24 x 4, then enter the data. The calculator enters data into a matrix by going sequentially across the rows. So the first row will be 10 10 10 3.32, and the second row will be 6 8 5 2.26. After the data is entered into matrix [A] press 2nd MODE to exit the Matrix editor and then execute the **MULREG** program.

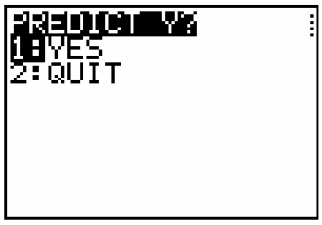










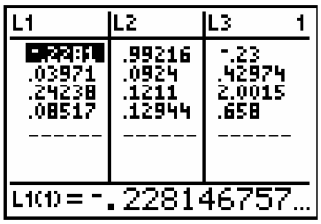


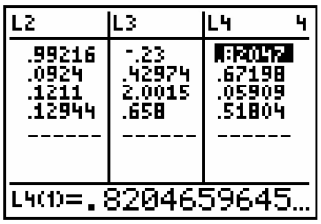
The program first reminds you that it is using matrix A as input. It then reminds you where the coefficients and their associated statistics are. Press ENTER to continue. Next, we see the standard deviation of the residuals, R^2 and R^2 adjusted (a penalized R^2 for having predictors in the model that are not useful), and the degrees of freedom for the ANOVA table. Press ENTER to continue. Next we see the sums of squares for the ANOVA table, the F statistic and its p-value. Press ENTER to continue. We are now asked if we want to use the model to predict a response given a set of predictors. For now, press 2 to quit.

The displayed ANOVA table gives the degrees of freedom (DF), the sum of squares (SS), and the mean square (MS) for each of the regression model, error, and total. The squared correlation coefficient, the F-statistic and p-value of the F-test, and the standard deviation S are also displayed. The value of R^2 here is telling us that only 32.8% of the variation in GPAs is explained by the high school variables.

The null hypothesis for the F-test is that each variable coefficient in the linear regression model is equal to 0 or H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. In this case, the p-value is 0.0433, which is a value usually considered low enough to be statistically significant. Thus, even with this small sample of size 24, we have enough evidence to reject H_0.

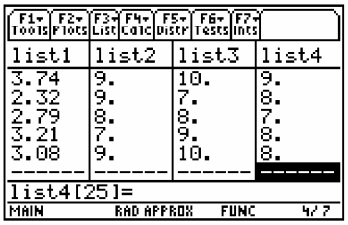
The program stores the computed regression coefficients and their associated statistics in L1 through L4. In this case, our model becomes μ_{GPA} = –0.228 + 0.0397*HSM + 0.2424*HSS + 0.0852*HSE. The Standard errors are in L2, with the associated *t* statistics in L3. Finally, the *p*-values for testing individually H_0: $\beta_i = 0$ are in L4. In this case, the only coefficient statistically different from 0 is for the second predictor, high school Science.

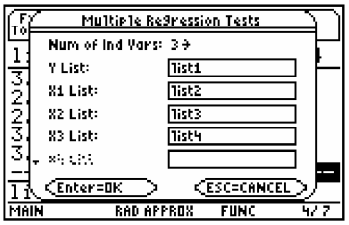


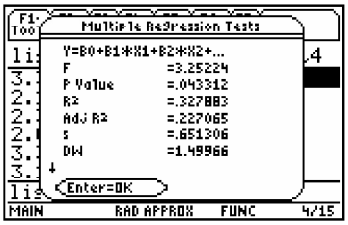


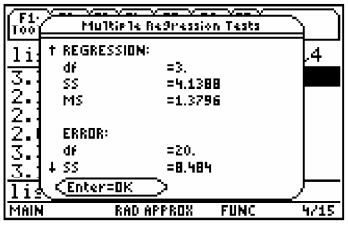
TI-89 Solution. Enter the data into four lists. From the [F6] Tests menu, select option B:MultRegTests. The first input item asks for the number of independent variables. Press the right arrow to expand the box. Here, select 3. Then use [VAR-LINK] to input the name of the list containing the dependent (*Y*) variable and the list names for the independent variables.

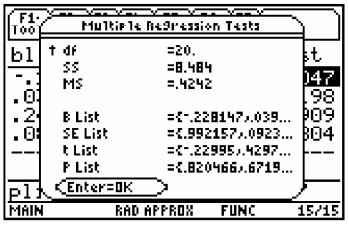
The first portion of the output shows the overall *F*-statistic for the model and its *p*-value, the multiple coefficient of determination (R^2) and the adjusted R^2 (this penalizes R^2 for adding predictor variables into a model that aren't helpful), the standard deviation of the residuals around the regression model and the Durbin-Watson statistic (a measure of autocorrelation among the residuals — since order of entry in this model has no real meaning, we ignore this here). Scrolling down the output, we find the ANOVA table for the model. We also see the first part of B list (list of coefficients for the model), SE list (the standard errors of the coefficients), T list (list of *t* statistics for each coefficient), and P list (list of *p*-values for each *t* statistic). Press [ENTER] to see these lists have been added into the List Editor. Scrolling left through the List Editor, we also see that a list of predicted values for each observation (yhat1), residuals (resid), standardized residuals (sresid), and leverage (lever), a measure of the influence of each observation on the regression, have also been added.

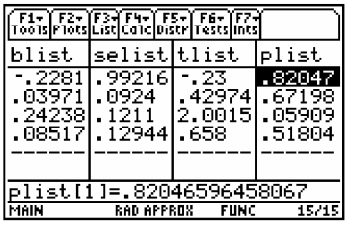






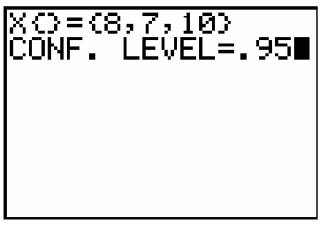


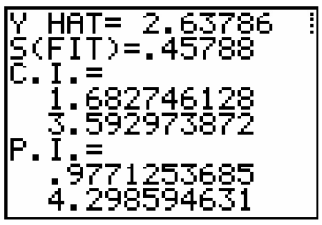




Example 11.2 More Grade Predictions. Use the preceding regression model to predict the GPA of a student with HSM = 8, HSS = 7, and HSE = 10.

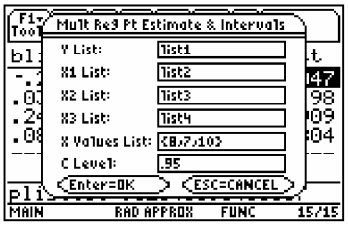
TI-83/84 Solution. Rerun program MULREG, but this time press [ENTER] to answer YES to the `Predict Y` question. You are next prompted to enter the X values. Enter them in curly braces ([2nd][(] and [2nd][)]) separated by commas. Press [ENTER]. You are now asked for the confidence level. Enter it as a decimal. Press [ENTER]. The next screen gives the actual prediction from the model along with the standard error of that prediction, the confidence interval, and the prediction interval.

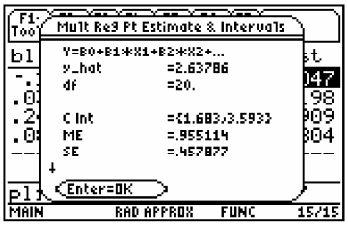


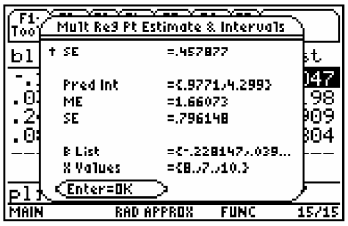


We see that the model predicts 2.64 as the GPA for students with these high school grades. We are 95% confident the *average* GPA for *all* students with these grades will be between 1.68 and 3.59. For an *individual* student with these grades, we would predict a GPA between 0.98 and 4.30 (4.0) with 95% confidence. The intervals are wide due to the low r^2 for the model.

TI-89 Solution. From the [F7] `Ints` menu, select option `8:MultRegInt`. You will first be asked for the number of independent variables (which should already be set to 3 from the prior regression). The list of values to predict for (these must be entered within curly braces which are [2nd][(] or [2nd][)]) and the confidence level. Pressing [ENTER], we see the predicted value from the equation, a confidence interval for the average GPA of all students with this combination of high school grades, a prediction interval for a particular student with these grades, and the beginning of the coefficient list.

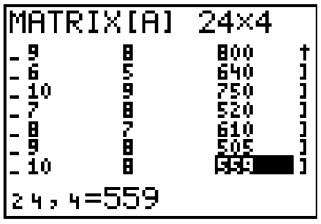


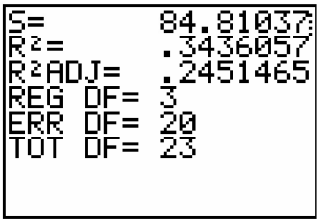


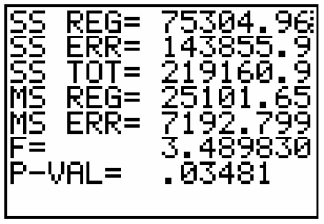


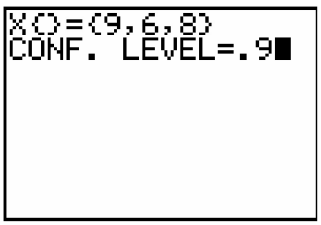
Example 11.3 Predicting SAT Math Scores. With the above 24 sample points, perform a regression analysis to predict SATM from the three high school grade variables. Then find the predicted SATM for a student with HSM = 9, HSS = 6, and HSE = 8, using 90% confidence.

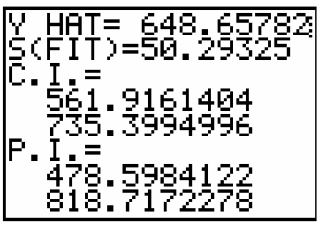
Solution. First, we must edit the last column of matrix [A] so that it contains the SATM scores. To enter these easily, use the down arrow after each entry rather than pressing ENTER. Then rerun the **MULREG** program to obtain a new regression model.

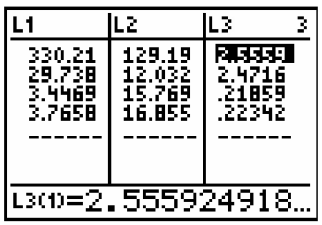












The new model is μ_{SATM} = 330.21 + 29.738*HSM + 3.4469*HSS + 3.7658*HSE. The displayed *p*-value of 0.0348 gives us statistical evidence to reject that all regression coefficients of HSM, HSS, and HSE are 0. Thus, at least one coefficient is nonzero; observing the *t* statistics in L3, the only one that is large is for the first predictor variable (high school Math grade). The model predicts a SATM score for a student with these high school grades of about 649 (practically 650); however, the standard error of this prediction is large.

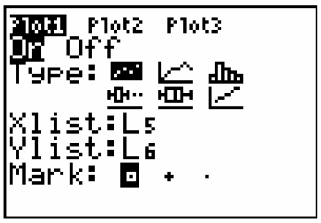
11.2 A Case Study

Example 11.4 On-line Brokerages. Below are some data for the top ten internet brokerages. The variables are the market share of the firm, the number of internet accounts in thousands, and the total assets of the firm in billions of dollars.

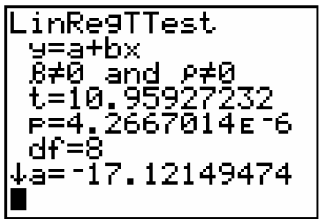
ID	Broker	Mshare	Accounts	Assets
1	Charles Schwab	27.5	2500	219.0
2	E*Trade	12.9	909	21.1
3	TD Waterhouse	11.6	615	38.8
4	Datek	10.0	205	5.5
5	Fidelity	9.3	2300	160.0
6	Ameritrade	8.4	428	19.5
7	DLJ Direct	3.6	590	11.2
8	Discover	2.8	134	5.9
9	Suretrade	2.2	130	1.3
10	National Discount Brokers	1.3	125	6.8

(a) Use a simple linear regression to predict assets using the number of accounts. Give the regression equation and the results of the significance test for the regression coefficient.
(b) Do the same using market share to predict assets.
(c) Run a multiple regression using both the number of accounts and market share to predict assets. Give the multiple regression equation and the results of the significance test for the two regression coefficients.
(d) Compare the results of parts (a), (b), and (c).

Solution. (a) We will enter the data into lists L4, L5, and L6. Then we define a scatterplot using Assets (in L6) on the *Y* axis and Accounts (in L5) on the *X* axis. The plot is very roughly linear; Fidelity and Charles Schwab clearly will influence the regression. We shall use LinRegTTest from the STAT TESTS menu. Then we execute the LinRegTTest on lists L5 and L6 to obtain a linear regression equation predicting assets using the number of accounts.

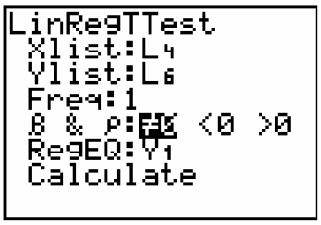


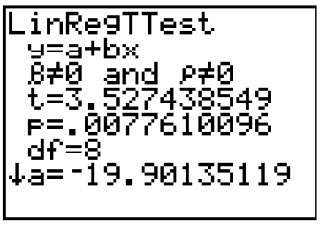


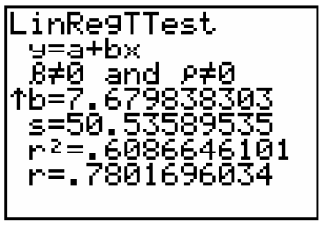


The *p*-value for a two-sided significance test is 4.2667×10^{-6}. If the regression coefficient β_1 were equal to 0, then there would be only a very small chance of obtaining a value of *b* as large as 0.0832, or a correlation as high as 0.96827, even with such a small sample. Thus, we have strong evidence to reject the hypothesis that $\beta_1 = 0$.

(b) Now we execute the LinRegTTest on lists L4 and L6 to obtain a linear regression equation $Assets = a + bx = -19.901 + 7.680\,MarketShare$.



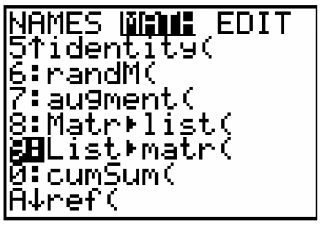




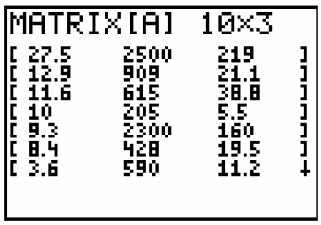
The *p*-value for the two-sided significance test is 0.007761. If the regression coefficient β_1 were equal to 0, then there would be less than a 1% chance of obtaining a value of *b* as high as 7.68 or a correlation as high as 0.78. Thus, we again can reject the hypothesis that $\beta_1 = 0$.

(c) To execute the **MULREG** program, we must first move the data to matrix [A]. To do so, we can use the List▶matr(command from the MATRX MATH menu. After

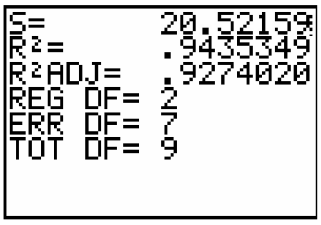
bringing this command to the Home screen, enter the command List▶matr(L4,L5,L6,[A]), where [A] is retrieved from the MATRIX NAMES menu.

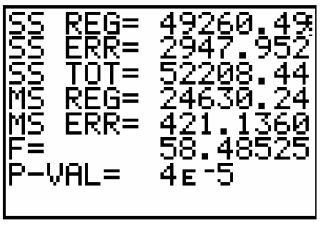


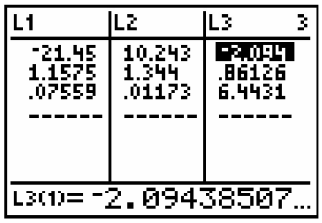




Last, execute the **MULREG** program.





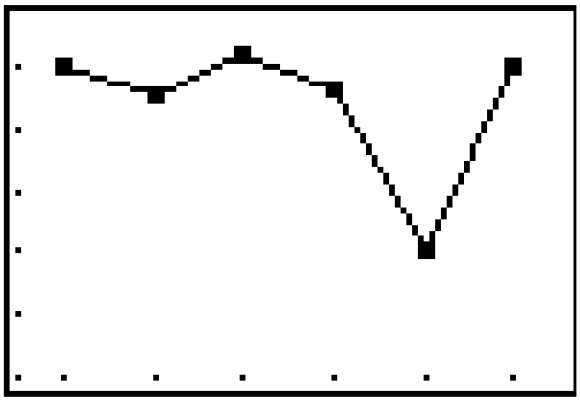


The regression equation is $\mu_{ASSETS} = -21.45 + 1.1575\,MarketShare + 0.0756\,Accts$. The p-value of (approximately) 0 gives significant evidence to reject H_0: $\beta_1 = \beta_2 = 0$.

(d) Comparing the r^2 values from each part, we see that the multiple linear regression r^2 of 94.35% is not really different from the 93.8% obtained from the simple model using only the number of accounts. It seems reasonable to believe that market share (which has a very small t statistic) and number of accounts might be collinear (related to each other). The simple regression using only the number of accounts is most likely a "best" model.

CHAPTER

12



One-Way Analysis of Variance

12.1	Inference for One-Way Analysis of Variance
12.2	Comparing the Means

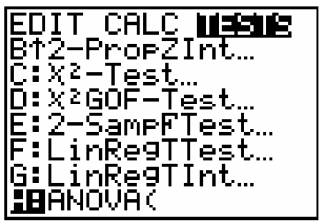
Introduction

Just as a chi-squared test can be seen as an extension of the two-proportion test, One-way Analysis of Variance (ANOVA) is an extension of the independent samples *t*-test to more than two groups.

In this chapter, we perform one-way analysis of variance (ANOVA) to test whether several normal populations, assumed to have the same variance, also have the same mean. The procedure is so similar on all models, that no additional explanation for the TI-89 is necessary.

12.1 Inference for One-Way Analysis of Variance

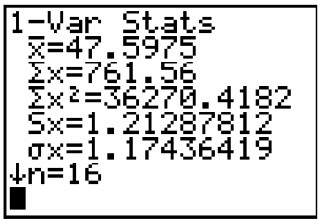
We begin with an exercise that demonstrates built-in analysis of variance capabilities using the ANOVA(command from the STAT TESTS menu.

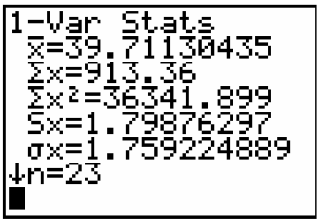


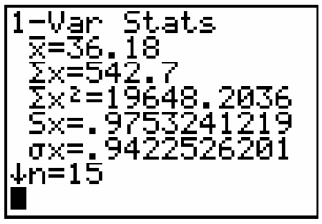
Example 12.1 Comparing Tropical Flowers. The data below give the lengths in millimeters of three varieties of the tropical flower *Heliconia*, which are fertilized by different species of hummingbird on the island of Dominica. Perform an ANOVA test to compare the mean lengths of the flowers for the three species.

H. bihai							
47.12	46.75	46.81	47.12	46.67	47.43	46.44	46.64
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36
H. caribaea red							
41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57
39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07
38.10	37.97	38.79	38.23	38.87	37.78	38.01	
H. caribaea yellow							
36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.1
35.17	36.82	36.66	35.68	36.03	34.57	34.63	

Solution. The ANOVA(command requires the data to be in lists. Here we shall enter the data into lists L1, L2, and L3. After doing so, we evaluate the basic statistics of each list.



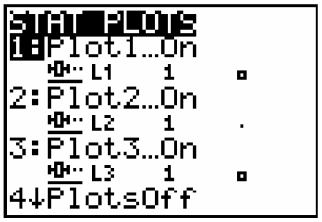


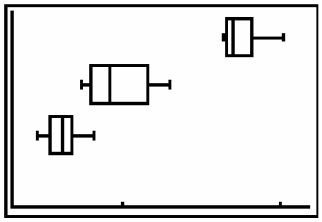


These sample means range from slightly larger than 36 to more than 47.5. Is the difference real (statistically significant) or just due to randomness?

Before proceeding to the ANOVA, we must consider the assumptions of the test. The first is that all groups have the same *population* standard deviation. In practice, the results will be approximately correct if the largest standard deviation is no more than twice as large as the smallest. Since 2*.975 = 1.95 > 1.799, we are safe to continue based on this criterion.

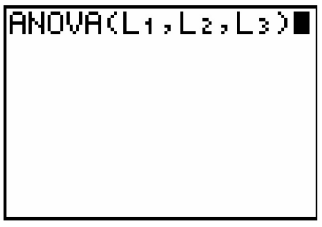
The second assumption is that the data come from Normal populations. We could do a Normal plot for each type of flower, but in this case side-by-side boxplots are a good idea. We're looking for indications of skewness or outliers. If you've forgotten how to make a boxplot, refer to Chapter 1 of this manual. Once all three plots are defined, press ZOOM 9 to display the graphs.

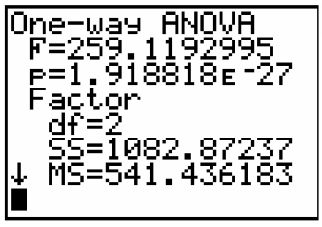


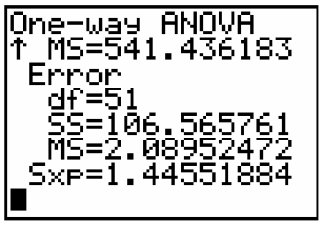


These data are not perfectly symmetric around the medians; also, there may be some skew in the distribution of *H. bihai*. However, with no outliers, we'll rely on the robustness of the ANOVA *F* test in continuing, as we have reasonable sample sizes.

Next, we test the hypothesis that the mean lengths of the three species are equal: $H_0 : \mu_1 = \mu_2 = \mu_3$. To do so, we bring the ANOVA(command to the Home screen and enter the command ANOVA(L1,L2,L3).



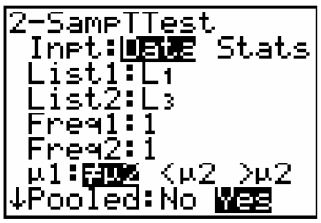


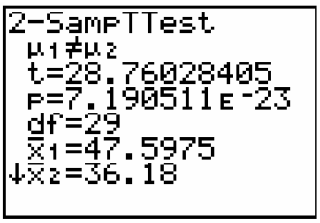


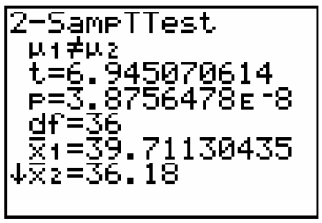
We receive a *p*-value of 1.92×10^{-27} from an *F*-statistic of 259.1193. The pooled standard deviation value is also displayed as Sxp=1.4455. If the true means were equal, then there would be almost no chance of the sample means varying by as much as they do with samples of these sizes. Thus, we have significant evidence to reject the claim that the mean lengths of these species are equal. We note that the r^2 value is not displayed, but it can be computed from the two displayed SS values. Here we can use r^2 = SSF/(SSF + SSE) = 1082.87237/(1082.87237 + 106.565761) = 0.9104.

Because we have rejected the hypothesis that the mean lengths are all equal, we can say that there is at least one pair of species that have different means. From the summary statistics, it appears that the species *H. bihai* and *H. caribaea yellow* have different mean lengths. But the sample means of *H. caribaea red* and *H. caribaea yellow* are close enough so that one might hypothesize that these species have the same mean length.

We can test any pair of species for equality of mean very quickly using 2-SampTTest, but need to pay attention to *p*-values that might cause problems due to multiple comparisons. Since one of the assumptions of ANOVA is that the population standard deviations are the same we will answer Yes to Pooled. We demonstrate below with the pairs (L1, L3) and (L2, L3). Due to the extremely low *p*-values (7×10^{-23} for L1 and L3 and 4×10^{-8} for L2 and L3), we see that we can reject both $\mu_1 = \mu_3$ and $\mu_2 = \mu_3$.







ANOVA for Summary Statistics

When raw data are given, then we can enter the data into lists and use the built-in ANOVA(command to test for equality of means. However, sometimes the summary statistics are given instead. In this case, we can use the **ANOVA1** program below to perform the analysis of variance on a TI-83/84 calculator; this is a built-in capability on a TI-89 — simply tell the calculator the data method is Stats and you will be prompted for the sample statistics instead of list names.

The ANOVA1 Program

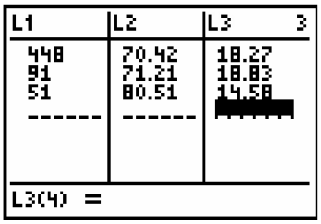
```
PROGRAM:ANOVA1
:1-Var Stats L2,L1
:sum(seq(L1(I)(L2(I)-x̄)^2,I,1,dim(L1)))→A
:sum(seq((L1(I)-1)(L3(I)^2),I,1,dim(L1)))→B
:dim(L/)→I
:(A/(I-1))/(B/(n-I))→F
:1-Fcdf(0,F,I-1,n-I)→P
:√(sum(seq((L/(J)-1)*L‹(J)^2,J,1,dim(L1)))
 /(n-I))→S
:ClrHome
:Output(2,1,"MEAN")
:Output(2,6,x̄)
:Output(3,1,"SP")
:Output(3,6,S)
:Output(4,1,"MSG")
:Output(4,6,A/(I-1))
:Output(5,1,"MSE")
:Output(5,6,B/(n-I))
:Output(6,1,"R^2")
:Output(6,6,A/(A+B))
:Output(7,1,"F")
:Output(7,6,F)
:Output(8,1,"P")
:Output(8,6,round(P,8))
```

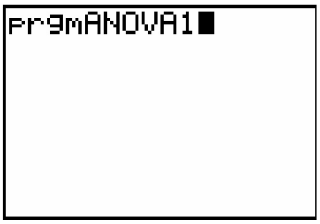
The program computes and displays the overall sample mean, the pooled sample deviation s_p, the mean square for groups MSG, the mean square for error MSE, the coefficient of determination, the *F*-statistic, and the *p*-value. Before executing the program, enter the sample sizes into list L1, the sample means into list L2, and the sample deviations into list L3.

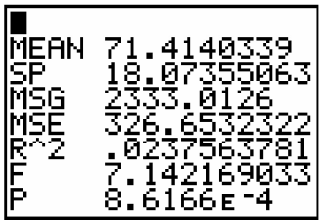
Example 12.2 Workplace Safety. The table below gives the summary data on safety climate index (SCI) as rated by workers during a study on workplace safety. Apply the ANOVA test and give the pooled deviation, the value of r^2, the F-statistic, and the p-value.

Job category	n	$\bar{x}$	s
Unskilled workers	448	70.42	18.27
Skilled workers	91	71.21	18.83
Supervisors	51	80.51	14.58

Solution. Enter the summary statistics into lists L1, L2, and L3, and execute the **ANOVA1** program. The desired statistics are then displayed. In particular, we obtain a pooled deviation of $s_p = 18.07355$ and a coefficient of determination of $r^2 = 0.023756$.







Due to the low p-value of 0.00086, we see that there is strong evidence to reject the claim that each group of workers has the same mean SCI. Looking at the given statistics, it is clear that supervisors rate the safety climate higher than the other two types of workers.

12.2 Comparing the Means

In this section, we provide a program for analyzing population contrasts based on summary statistics. Before executing the **CONTRAST** program, enter the sample sizes into list L1, the sample means into list L2, the sample deviations into list L3 (these are the same as for program **ANOVA1**), and the contrast equation coefficients into list L4. When prompted during the program, enter either **1** or **2** for a one-sided or two-sided alternative. The program then displays the p-value for a significance test and a confidence interval for mean contrast.

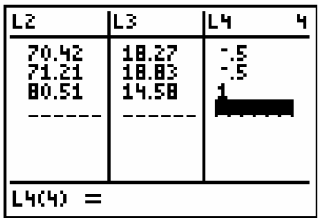
The CONTRAST Program

```
PROGRAM:CONTRAST
:Disp "ALTERNATIVE"
:Input A
:Disp "CONF. LEVEL"
:Input R
:1-Var Stats L2,L1
:√(sum(seq((L1(I)-1)*L3(I)^2,
 I,1,dim(L1)))/(n-dim(L1)))→S
:sum(seq(L2(I)L4(I),I,1,dim(L1)))→C
:S√ (sum(seq(L4(I)^2/L1(I),I,1,
 dim(L1))))→E
:C/E→T
:1-Var Stats L1
:Σx-dim(L1)→D
:0.5-tcdf(0,abs(T),D)→P
:ClrHome
:Disp "P-VALUE"
:If A=1
:Then
:Disp P
:Else
:Disp 2*P
:End
:"tcdf(0,X,D)"→Y1
:solve(Y1-R/2,X,2)→Q
:Disp "CONF.INTERVAL"
:Disp {round(C-QE,3),
 round(C+QE,3)}
```

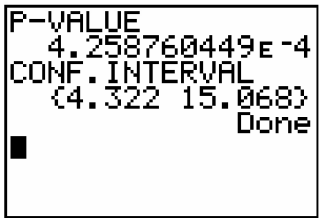
Example 12.3 More Workplace Safety. Using the data from Example 12.2 above, perform a two-sided significance test for the contrast between the average SCI of supervisors compared with the other two groups of workers. Also, find a 95% confidence interval for the contrast.

Solution. We use the contrast equation $\Psi = -.5\mu_{UN} - .5\mu_{SK} + 1\mu_{SU}$, and test the hypothesis H_0: $\mu_{SU} = .5\mu_{UN} + .5\mu_{SK}$ with the alternative H_A: $\mu_{SU} \neq .5\mu_{UN} + .5\mu_{SK}$.

If we still have the sample sizes, sample means, and sample deviations entered into lists L1, L2, and L3, then we just need to add the coefficients of Ψ into list L4. Then we execute the **CONTRAST** program by entering **2** for two-sided alternative and entering the desired confidence level.



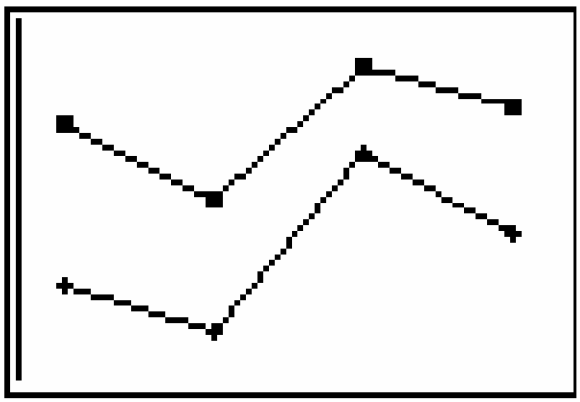




With the low *p*-value of 0.00043, we see that there is a significant contrast. According to the displayed confidence interval, the mean SCI rating by supervisors could be from 4.322 to 15.068 points higher than the average of the mean SCI ratings by unskilled workers and skilled workers.

CHAPTER

13



Two-Way Analysis of Variance

13.1	Plotting Means
13.2	Inference for Two-Way ANOVA

Introduction

Two-way analysis of variance is really best performed with a full statistical package on a computer. However, with a fairly lengthy program, TI-83 and -84 calculators can do a good job of this. TI-89 calculators easily generate the ANOVA table.

In this chapter, we provide a program that performs two-way analysis of variance to test for equality of means simultaneously among populations and traits in a two-factor experiment, and detail how a TI-89 might be used.

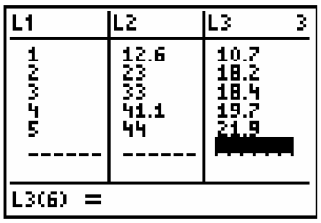
13.1 Plotting Means

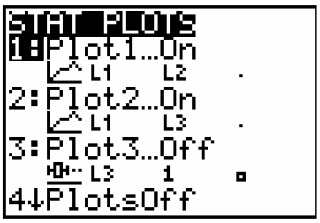
Example 13.1 Time Spent Eating. The table below gives the mean length of time (in minutes) that various groups of people spent eating lunch in various settings. Plot the group means for this example.

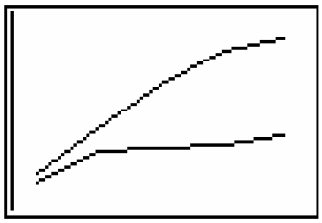
	Number of people eating				
Lunch setting	1	2	3	4	5
Workplace	12.6	23.0	33.0	41.1	44.0
Fast-food restaurant	10.7	18.2	18.4	19.7	21.9

Solution. We note that we cannot use ANOVA to analyze this data because each cell contains the *mean* of an unknown number of subjects, and we cannot assume that there are a fixed number of measurements per cell. In order to perform two-way ANOVA, we also would need to know the sample sizes and the variances of the data that produced the mean of each cell. However, we can plot the means with a time plot in order to observe a possible interaction.

First, enter the integers 1 through 5 into list L1, the five means for workplace diners into list L2, and the five means for the fast foods into list L3. Next set Plot1 for a time plot of L1 and L2, and set Plot2 for a time plot of L1 and L3.







We see that the patterns are not parallel, so it appears that we have an interaction.

13.2 Inference for Two-Way ANOVA

In this section, we give an **ANOVA2W** program that can be used on TI-83 and -84 calculators to perform two-way analysis of variance given the data. Two-way ANOVA for balanced designs (equal observations per cell) is a built-in function on the TI-89. This is a long program; we recommend it be downloaded from the text's website using a TI-Connect cable. Once your instructor (or a fellow student) has downloaded the program it is easily copied from one calculator to another (see Chapter 0 for more details on this).

The ANOVA2W Program

```
PROGRAM:ANOVA2W
Disp "DATA IN [A]"
Disp "","FACTOR A =
ROWS?","1=YES","2=NO",""
Repeat F≠0:getKey→F:End
If F=92:[A]^T→ [A]
Input "NO. B LEVELS=",B
dim([A]) →LθA: LθA (2) →A: LθA
(1)/B→N
{0}→LθB
{B,A}→dim([B])
For(I,1,B):For(J,1,A)
sum(seq([A]((I-1)N+K,J),K,1,N)) → [B](I,J)
End:End
For(J,1,A)
sum(seq([B](I,J),I,1,B)) →LθA(J)
End
For(I,1,B)
sum(seq([B](I,J),J,1,A)) →LθB(1)
End
sum(LθA)^2/(ABN) →C
Fοr(I,1,BN).Fοr(J,1,A)
Ans+[A](I,J) O
End:End
Ans-C→T
If F=92:[A]^T→ [A]
sum(LθA^2)/(BN)-C→U
sum(LθB^2)/(AN)-C→V
For(I,1,B):For(J,1,A)
Ans+[B](I,J)^2
End:End
Ans/N-U-V-C→W
If N>1
Menu("FIT INTERACTION?","YES",G,"NO",H)
Else
Goto H
End
Lbl G
T-U-V-W→Z
ClrHome
Output(1,1,"SSA="
Output(1,9,U
Output(2,1,"SSB="
Output(2,9,V
Output(3,1,"SSAB="
Output(3,9,W
Output(4,1,"SSE="
Output(4,9,Z
Output(5,1,"SSTO="
Output(5,1,"       "
Pause
ClrHome
Output(1,1,"F A="
Output(1,9,U/Z
Output(2,1," p A="
1-Fcdf(0,U/Z,A,K) →P
round(P,6) →P
Output(2,9,P
Output(3,1,"F B="
Output(3,9,V/Z
Output(4,1," p B="
1-Fcdf(0,V/Z,B,K) →P
round(P,6) →P
Output(4,9,P
Output(5,1,"F AB="
Output(5,9,W/Z
Output(6,1," p AB="
1-Fcdf(0,W/Z,J,K) →P
round(P,6) →P
Output(6,9,P
Output(7,1,"       "
Pause
Goto R
Lbl H
ClrHome
T-U-V→Z
Output(1,1,"ASSUME NO"
Output(2,1,"INTERACTION"
Output(3,1,"SSA="
Output(3,9,U
Output(4,1,"SSB="
Output(4,9,V
Output(5,1,"SSE="
Output(5,9,Z
Output(6,1,"SST="
Output(6,9,T
Output(7,1,"    "
Pause
ClrHome
A-1→A
Output(1,1,"DF A="
Output(1,9,A
U/A→U
B-1→B
Output(2,1,"DF B="
Output(2,9,B
V/B→V
(A+1)(B+1)N-A-B-1→K
Output(3,1,"DF E="
```

```
Output(5,9,T
Output(6,1,"     "
Pause
ClrHome
A-1→A
Output(1,1,"DF A="
Output(1,9,A
U/A→U
B-1→B
Output(2,1,"DF B="
Output(2,9,B
V/B→V
(A)(B) →J
Output(3,1,"DF AB="
Output(3,9,J
W/J→W
(A+1)(B+1)(N-1) →K
Output(4,1,"DF E="
Output(4,9,K
Z/K→Z
Output(5,1,"DF T="
(A+1)(B+1)N-1→G
Output(5,9,G
Pause
ClrHome
Output(1,1,"MSA="
Output(1,9,U
Output(2,1,"MSB="
Output(2,9,V
Output(3,1,"MSAB="
Output(3,9,W
Output(4,1,"MSE="
Output(4,9,Z
Output(3,9,K
Z/K→Z
Output(4,1,"DF T="
Output(4,9,(A+1)(B+1)N-1
Output(5,1,"MSA="
Output(5,9,U
Output(6,1,"MSB="
Output(6,9,V
Output(7,1,"MSE="
round(Z,6) →Z
Output(7,9,Z
Pause
ClrHome
Output(1,1,"F A="
Output(1,9,U/Z
Output(2,1," p A="
1-Fcdf(0,U/Z,A,K) →P
round(P,6) →P
Output(2,9,P
Output(3,1,"F B="
Output(3,9,V/Z
Output(4,1," p B="
1-Fcdf(0,V/Z,B,K) →P
round(P,6) →P
Output(4,9,P
Pause
Lbl R
ClrHome
1/N[B] → [B]
Output(1,1,"MEANS IN MATRX B"
Output(2,1,"COLS FACTOR A"
Pause
ClrHome
Stop
```

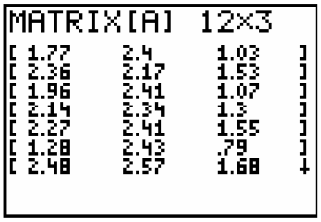
To execute the program, enter the data into matrix [A]. The program displays the *F*-statistics and *p*-values for the two factors, and for the interaction when there is more than one observation per cell (or interaction was not specified in the model).

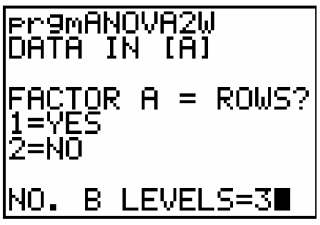
Example 13.2 Iron in Food. Does the type of cooking pot affect the iron content in food, and does the type of food cooked matter? In many parts of the world where people suffer from anemia, this could be important information. The table below gives the amount of iron in certain foods, measured in milligrams of iron per 100 grams of cooked food, after samples of each food were cooked in each type of pot.

IRON					**Food**							
Type of pot	Meat				Legumes				Vegetables			
Aluminum	1.77	2.36	1.96	2.14	2.40	2.17	2.41	2.34	1.03	1.53	1.07	1.30
Clay	2.27	1.28	2.48	2.68	2.41	2.43	2.57	2.48	1.55	0.79	1.68	1.82
Iron	5.27	5.17	4.06	4.22	3.69	3.43	3.84	3.72	2.45	2.99	2.80	2.92

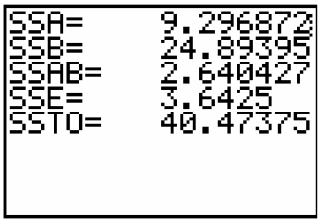
Perform two-way ANOVA on the data regarding the main effects and interaction, then plot the means and examine the plot.

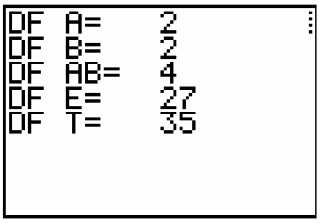
TI-83/84 Solution. Enter the data into matrix [A]. (Hint: it may sometimes be easier to enter data into lists and then use `List▶Matrix` from the `List Math` menu to create matrix [A].) It makes no difference what is considered factor A or B; here the data have been entered so that all meat observations are in the first column, legumes in the second, and vegetables in the third. All the aluminum pot observations come first, then clay pots, and finally iron pots. You are asked if factor A is in the rows. Since here we consider the type of food factor A, press 2 for No (otherwise the program will look for multiples of this across the rows). If you had entered the data just as it was in the table above (a 3 × 12 matrix) the answer here would be No. You are then asked for the number of levels of Factor B. Since there were three types of pots, there are 3 levels. Type 3 followed by ENTER. The next question is self-explanatory. One should always fit the interaction term unless prior examination shows this is not significant.

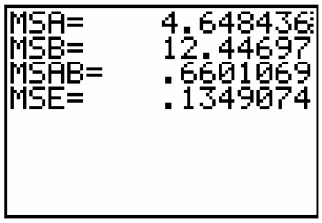


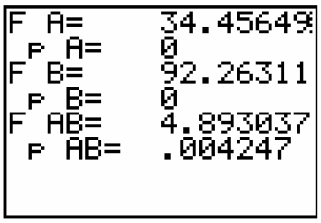


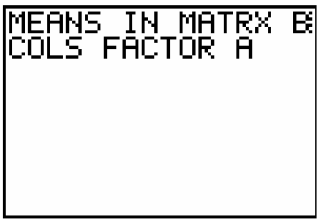


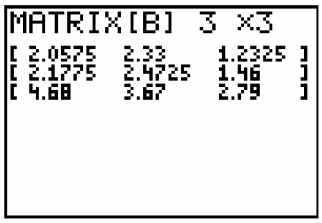










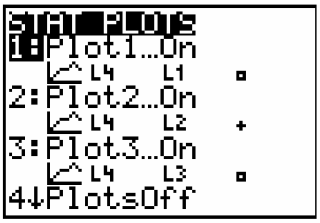


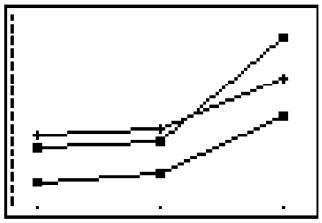
The program then displays the entries in the ANOVA table and pauses after each display. Press ENTER to proceed to the next. The program next reminds you that "cell" means are stored in matrix [B] and the columns of this matrix represent the levels of factor A (food here).

We obtain very low *p*-values for both factors and for the interaction. Therefore, we can conclude that, even with these small samples, there is a significant difference in the mean due to type of pot as well as type of food. Also, there is a statistically significant interaction.

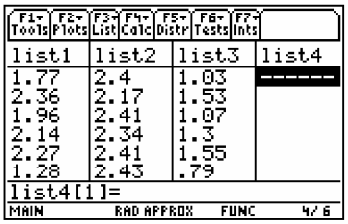
To create the means plot, use `Matr▸List` from the `Matrix Math` menu to create lists to plot. You need to specify as many lists as there are columns in [B]. Next, input a column containing values 1, 2, and 3 in `L4`. Define connected plots (time plots) using L4 as the `Xlist` and each list of means as the `Ylist`; in this example you will have three plots. Pressing ZOOM 9 displays the plots. Here, we can clearly see the interaction since the lines cross. Vegetables have consistently the lowest amount of iron but cooking them in an iron pot increases the amount. Meat cooked in an iron pot has the largest amount of iron due to the interaction.

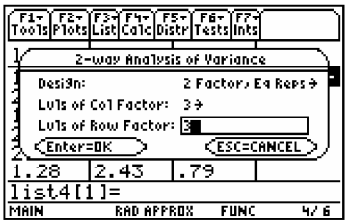


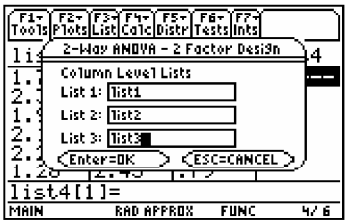


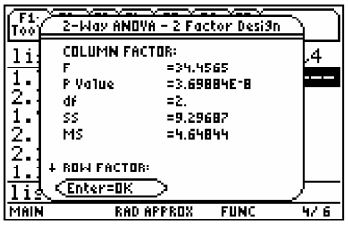


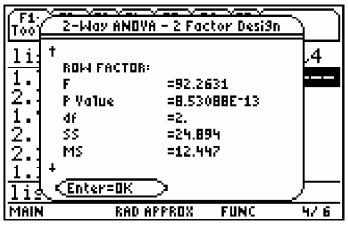
TI-89 Solution. Enter the data. Here, all meat observations are in `list1`, all legumes in `list2`, and all vegetables in `list3`. Select option `D:ANOVA2-Way` from the [F6] `Tests` menu. Use the right arrow key to expand the menu options on the first screen. We have a 2 Factor Equal Replications set of data with three levels of the column factor (the three lists) and three levels of the row factor. Press ENTER to continue. Next, indicate which lists contain your data. The calculator displays the entries of the ANOVA table.

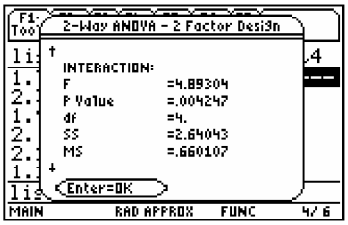


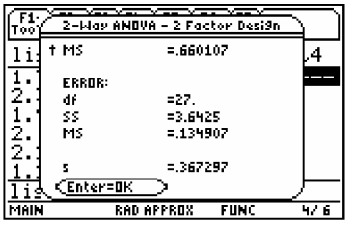




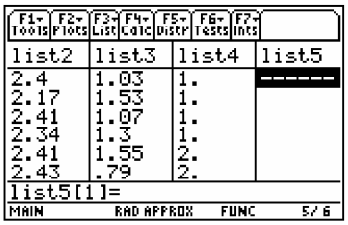


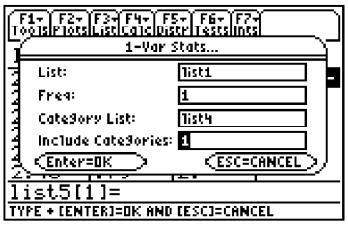


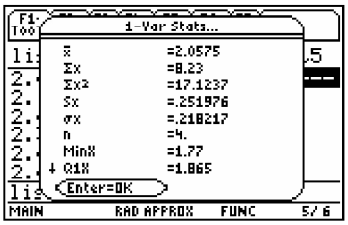




Unfortunately, the TI-89 does not create the cell means for you to create the means plot. You can calculate these fairly easily, though. Enter four 1s in list 4, then four 2s, then four 3 s. These will correspond to the type of pot (1 = Aluminum, 2 = Clay, 3 = Iron). Now, from the `Calc` menu, use `1-Var Stats` and the Category list to get statistics for each category for each list. Below, we see the mean of meat cooked in aluminum pots is 2.0575.

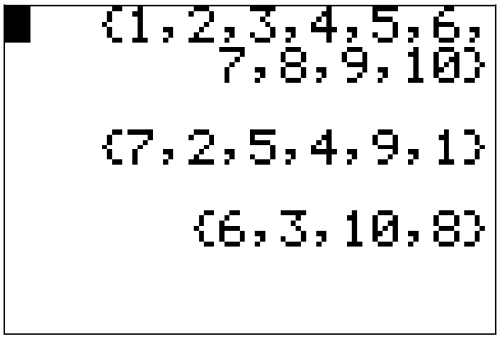






CHAPTER

14



Bootstrap Methods and Permutation Tests

14.1	The Bootstrap Idea
14.2	First Steps in Using the Bootstrap
14.3	How Accurate Is a Bootstrap Distribution?
14.4	Bootstrap Confidence Intervals
14.5	Significance Testing Using Permutation Tests

Introduction

In this chapter, we demonstrate the bootstrap methods for estimating population parameters. Throughout, we use TI-83 Plus programs to perform the necessary resampling procedures.

To truly perform these resampling methods correctly requires more processing power than a TI calculator possesses (we'd want several hundred resamples). The methods given in this chapter will approximate a resampling distribution only. Some of these programs will run for several minutes, depending on the size of the original sample and the number of resamples.

Bear in mind that since these methods are based on random resamples, your output will *not* duplicate ours.

14.1 The Bootstrap Idea

We first provide a TI calculator program that performs resampling on a previously obtained random sample. Before executing the **BOOT** program, enter the random sample into list L1. Then bring up the program and enter the desired number of resamples. If you also want a confidence interval for the statistic, enter **1** for CONF. INTERVAL?; otherwise, enter **0**. The program takes resamples from the entered random sample, enters their means into list L2, and displays the mean and bootstrap standard error of the resamples.

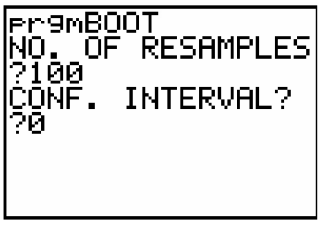
The BOOT Program

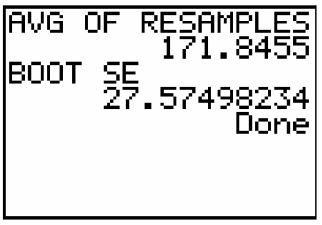
```
Program:BOOT
:dim(L1)→N
:Disp "NO. OF RESAMPLES"
:Input B
:Disp "CONF. INTERVAL?"
:Input Y
:If Y=1
:Then
:Disp "CONF. LEVEL"
:Input R
:"tcdf(0,X,N-1)" →Y1
:solve(Y1-R/2,X,2) →Q
:1-Var Stats L1
:x̄→X
:End
:ClrList L2
:For(I,1,B)
:randInt(1,N,N) →L3
:sum(seq(L1(L3(J)),J,1,N))/N→L2(I)
:End
:ClrList L3
:1-Var Stats L2
:ClrHome
:Disp "AVG OF RESAMPLES"
:Disp x̄
:Disp "BOOT SE"
:Disp Sx
:If Y=1
:Then
:Output(6,2,"CONF. INTERVAL")
:Output(7,2,X-Q*Sx)
:Output(8,2,X+Q*Sx)
:End
```

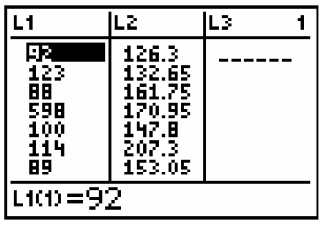
Example 14.1 Guinea Pig Survival. Here is an SRS of 20 guinea pig survival times (in days) after they were injected with *tubercle bacillus* during a medical trial. Create and inspect the bootstrap distribution of the sample mean for these data.

92	123	88	598	100	114	89	522	58	191
137	100	403	144	184	102	83	126	53	79

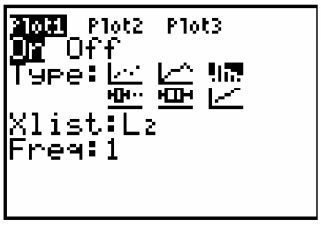
Solution. We will use the **BOOT** program with 100 resamples to create a bootstrap distribution. But first, we must enter the data into list L1. In this running of the program, we will not find a confidence interval, so we enter **0** for CONF. INTERVAL? when prompted. After taking a few minutes to execute, the program stores the 100 means from the resamples in list L2.

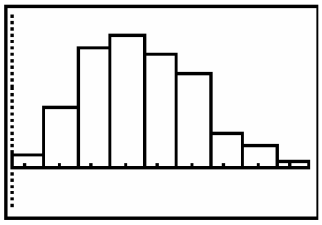






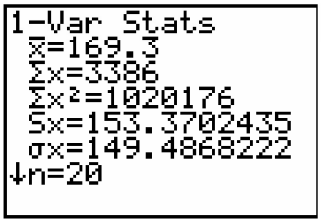
To assess the normality of the distribution of the resample means, we shall make a histogram of the data in list L2. We adjust the STAT PLOT settings and press ZOOM 9 to display the graph; if desired, the default settings can be adjusted using WINDOW. We see that the bootstrap distribution appears close to normal.

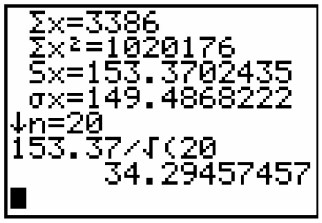




Example 14.2 Bootstrap Standard Error and the Original. Compute the standard error $s/\sqrt{n}$ for the 20 survival times in Example 14.1 and compare it with the bootstrap standard error from the resampling done in that exercise.

Solution. With the data entered into list L1, we can use the command 1-Var Stats L1 to compute the sample deviation *s*. We then divide Sx by $\sqrt{20}$ to compute the standard error.





We see that the standard error of 34.294629 is somewhat higher than the bootstrap standard error of 27.575 obtained from our 100 resamples in Example 14.1. This is due to the skew and outliers in the original distribution.

14.2 First Steps in Using the Bootstrap

In this section, we demonstrate how to compute a bootstrap *t* confidence interval for a population parameter. We begin though with a quick estimation of the bias obtained using the preceding exercises.

Example 14.3 More Guinea Pig Survival. Let's return to the 20 guinea pig survival times from Example 14.1.
(a) What is the bootstrap estimate of the bias?
(b) Give the 95% bootstrap *t* confidence interval for μ.
(c) Give the usual 95% one-sample *t* confidence interval.

Solution. (a) In Example 14.2, we found the sample mean of the data, along with the standard deviation, using the `1-Var Stats` command. From the output display, we see that $\bar{x}$ = 169.3. The average of all 100 resample means obtained in Example 14.1 was 171.846. Thus, the bootstrap estimate of bias is 171.846 – 169.3 = 2.546.

(b) We now apply the formula $\bar{x} \pm t^* SE_{BOOT}$, where $\bar{x}$ is the mean of the *original* sample, the critical value *t** is from the $t(n-1)$ distribution and the sample size is *n* = 20. To find this critical value from the *t*(19) distribution, we can use the **TSCORE** program (page 82), Table C, or `InvT` on a TI-84 or -89. Doing so, we obtain *t** = 2.093.

The 95% bootstrap *t* confidence interval for mean guinea pig survival (in days) after being injected with *tubercle bacillus* becomes $169.3 \pm 2.093 * 27.575$, or (111.586, 227.014).

Note: This confidence interval would have been computed and displayed when executing the **BOOT** program in Example 14.1 had we entered **1** for `CONF. INTERVAL?`. A different bootstrap confidence interval is displayed below from another execution of the **BOOT** program. Remember, since this is a different random resample; we do not expect the two to agree.

```
prgmBOOT
NO. OF RESAMPLES
?100
CONF. INTERVAL?
?1
CONF. LEVEL
?.95
```

```
AVG OF RESAMPLES
          173.868
BOOT SE
     39.40451412

 CONF. INTERVAL
 86.82540538
 251.7745946
```

(c) To compute the standard one-sample *t* confidence interval, we use the `TInterval` from the `STAT TESTS` menu. After adjusting the `List` to `L1`, we calculate a 95% confidence interval of (97.521, 241.08).

```
TInterval
 Inpt:Data Stats
 List:L1
 Freq:1
 C-Level:.95
 Calculate
```

```
TInterval
 (97.521,241.08)
 x̄=169.3
 Sx=153.3702435
 n=20
```

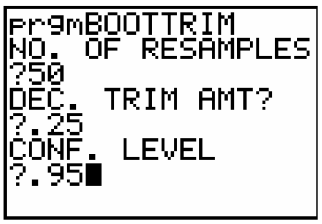
Confidence Interval for a Trimmed Mean

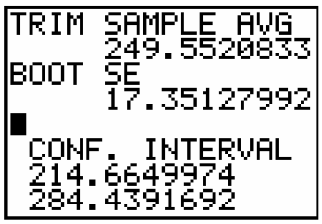
Next, we demonstrate a program that will compute a bootstrap *t* confidence interval for a trimmed mean (where x% of data at the top and bottom are deleted or trimmed, and a mean calculated). Before executing the **BOOTTRIM** program given on the next page, enter the random sample into list L1. Then enter the desired number of resamples, the decimal amount to be trimmed at each end, and the desired confidence level when prompted. The program takes resamples from the entered random sample and enters their trimmed mean into list L2. Then the trimmed mean of the original sample, the trimmed bootstrap standard error, and the confidence interval are displayed.

Example 14.4 Seattle Real Estate. The table below gives an SRS of 50 real estate sale prices in Seattle (in thousands of dollars) during 2002. Use the bootstrap *t* method to give a 95% confidence interval for the 25% trimmed mean sale price.

142	175	197.5	149.4	705	232	50	146.5	155	1850
132.5	215	116.7	244.9	290	200	260	449.9	66.407	164.95
362	307	266	166	375	244.95	210.95	265	296	335
335	1370	256	148.5	987.5	324.5	215.5	684.5	270	330
222	179.8	257	252.95	149.95	225	217	570	507	190

Solution. After entering the data into list L1, we simply execute the **BOOTTRIM** program. Here we used 50 resamples. The trimmed mean of the original sample is given as 249.552, and the 25% trimmed bootstrap confidence interval for the mean price of a house in Seattle in 2002 is 214.665 to 284.439 thousand dollars.





Note: Due to programming differences, the trimmed mean given by the **BOOTTRIM** program may not agree with the value given by software. Because the sample size of 50 is not a multiple of 4, we cannot trim precisely 25% of the measurements from the high and low ends. In the case of Example 14.4 above, the **BOOTTRIM** program trimmed the lowest 12 and the highest 13 measurements before computing the trimmed sample mean.

The BOOTTRIM Program

```
Program:BOOTTRIM
:dim(L1)→N
:Disp "NO. OF RESAMPLES"
:Input B
:Disp "DEC. TRIM AMT?
:Input M
:Disp "CONF. LEVEL"
:Input R
:"tcdf(0,X,N-1)" →Y1
:solve(Y1-R/2,X,2) →Q
:int(M*N)+1→L
:int((1-M)*N) →U
:SortA(L1)
:sum(seq(L1(I),I,L,U))/(U+1-L) →X
:ClrList L2,L4
:For(I,1,B)
:randInt(1,N,N) →L3
:For(J,1,N)
:L1(L3(J)) →L4(J)
:End
:SortA(L4)
:sum(seq(L4(I),I,L,U))/(U+1-L) →L2(I)
:End
:ClrList L3,L4
:1-Var Stats L2
:ClrHome
:Disp "TRIM SAMPLE AVG"
:Disp X
:Disp "BOOT SE"
:Disp Sx
:Output(6,2,"CONF. INTERVAL")
:Output(7,2,X-Q*Sx)
:Output(8,2,X+Q*Sx)
:Stop
```

Difference in Means

We conclude this section with a program that computes a bootstrap *t* confidence interval for the difference in means. Before executing the **BOOTPAIR** program, enter a random sample from the first population into list L1 and enter a random sample from the second population into list L2. Then enter the desired number of resamples and the confidence level. The resampled differences in mean are stored in list L3. The difference of the original sample averages is displayed along with the bootstrap standard error and the confidence interval.

The BOOTPAIR Program

```
Program:BOOTPAIR
:dim(L1)→N
:dim(L2)→M
:Disp "NO. OF RESAMPLES"
:Input B
:Disp "CONF. LEVEL"
:Input R
:1-Var Stats L1
:x̄→X
:1-Var Stats L2
:x̄→Y
:ClrList L3,L4
:For(I,1,B)
:randInt(1,N,N)→L5
:randInt(1,M,M)→L6
:sum(seq(L1(L5(J)),J,1,N))/N→L3(I)
:sum(seq(L2(L6(J)),J,1,M))/M→L4(I)
:End
:1-Var Stats L3
:Sx→S
:1-Var Stats L4
:Sx→T
:int((S²/N+T²/M)²/((S²/N)²/
(N-1)+(T²M)²/(M-1)))→P
:√(S²/N+T²/M)→D
:"tcdf(0,X,P)"→Y1
:solve(Y1-R/2,X,2)→B
:B*D→E
:ClrHome
:Disp "DIFF OF AVGS"
:Disp X-Y
:Disp "BOOT SE"
:Disp D
:Output(6,2,"CONF. INTERVAL")
:Output(7,2,X-Y-E)
:Output(8,2,X-Y+E)
```

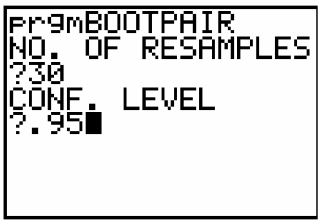
Example 14.5 Reading Scores. Following are the scores on a test of reading ability for two groups of third-grade students. The Treatment group used a set of "directed reading activitites" and the Control group followed the same curriculum without the activities.

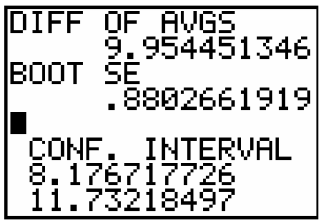
Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

(a) Bootstrap the difference in means $\bar{x}_1 - \bar{x}_2$ and give the bootstrap standard error and a 95% bootstrap *t* confidence interval.
(b) Compare the bootstrap results with a two-sample *t* confidence interval.

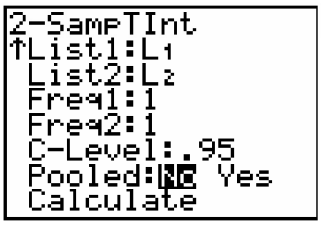
Solution. (a) We first enter the treatment scores into list L1 and the control scores into list L2. We shall use 30 resamples to bootstrap the difference in means and to obtain a 95% confidence interval. Upon executing the **BOOTPAIR** program, we obtain a bootstrap standard error of about 0.880 and a 95% confidence interval for the difference

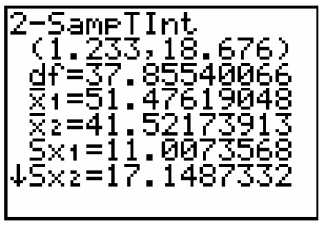
in mean reading scores for the two teaching methods of (8.177, 11.732). Since the interval does not contain 0, we have evidence that the directed reading activities help students reading ability.





(b) To find a traditional two-sample t confidence interval, we can use `2-SampTInt` from the `STAT TESTS` menu. Because we are not assuming normal populations with the same variance, we do not pool the sample variances. Upon calculating, we obtain a 95% confidence interval of (1.233, 18.676). This interval is much wider than the bootstrap interval due to the much larger standard error obtained when using the original sample deviations s_x and s_y as opposed to the sample deviations from the collection of resample means.



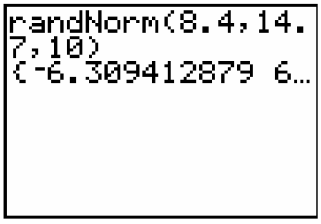


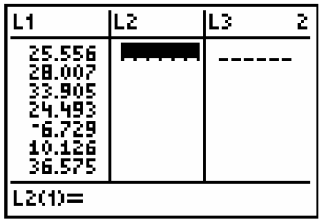
14.3 How Accurate Is a Bootstrap Distribution?

Below we work an exercise to demonstrate how the bootstrap distribution varies with the sample size.

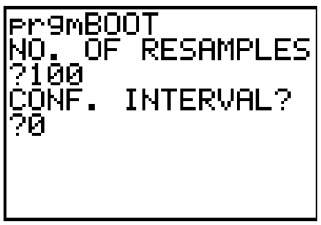
Example 14.6 Bootstrapping a Normal Distribution. Draw an SRS of size 10 from a $N(8.4, 14.7)$ population. What is the exact distribution of the sampling mean $\bar{x}$ for this sample size? Bootstrap the sample mean using 100 resamples. Give a histogram of the bootstrap distribution and the bootstrap standard error. Repeat the process for a sample of size 40.

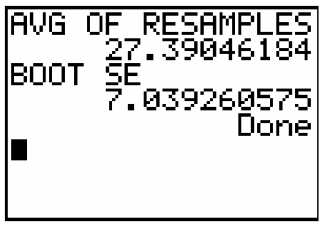
Solution. For an SRS of size n = 10 from a $N(8.4, 14.7)$ population, $\bar{x} \sim N(8.4, 14.7/\sqrt{10}) = N(8.4, 4.6845)$. To draw such an SRS, we will use the `randNorm(` command from the `MATH PRB` menu. Enter the command `randNorm(8.4,14.7,10)` `STO▸` `L1` to store the SRS in list `L1`.

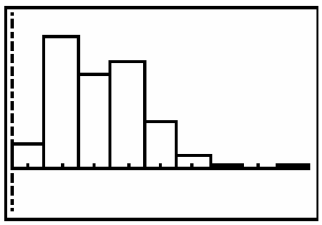




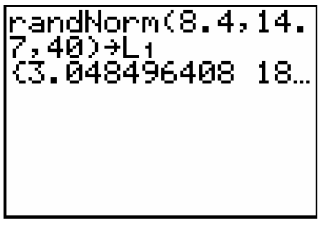
With the sample in list L1, we can execute the **BOOT** program to bootstrap the sample mean. In this case, we enter **0** for CONF. INTERVAL? because we are only interested here in the bootstrap distribution. Due to the small sample size, we obtain statistics of this bootstrap distribution that differ noticeably from those of the $N(8.4,4.6845)$ distribution. A histogram of the resample means stored in list L2 is shown; this distribution does not look Normally distributed.

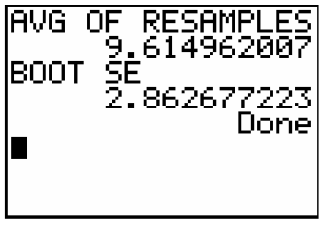


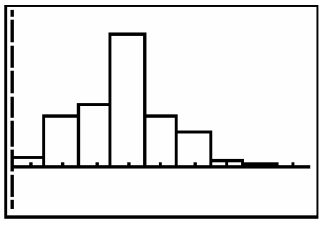




(c) We now repeat the process for an SRS of size 40. With this sample size, the bootstrap distribution statistics become much closer to the actual $N(8.4,14.7/\sqrt{40}) - N(8.4,2.324)$ distribution of $\bar{x}$.





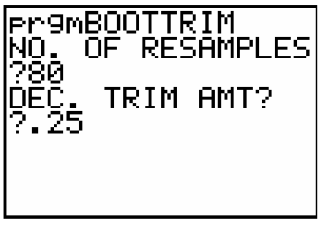


14.4 Bootstrap Confidence Intervals

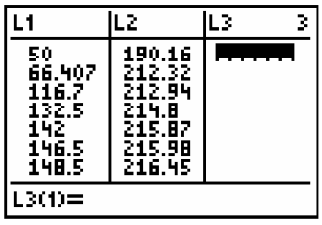
In this section, we demonstrate how to find a confidence interval based solely on the bootstrap distribution. We also provide another program that will bootstrap the correlation coefficient or the regression slope for two related variables.

Example 14.7 More Seattle Real Estate. The Seattle house prices in Example 14.4 are strongly right-skewed with outliers. In Example 14.4, we calculated a bootstrap interval for a 25% trimmed mean, using a *t* critical value. How does this interval compare to one based on the bootstrap distribution?

Solution. We run the **BOOTTRIM** program using 80 resamples since the endpoints of our interval will be the 2.5th and 97.5th percentiles of the bootstrap distribution. Since .025*80 = 2, the endpoints will be the 3rd and 78th observations in a sorted bootstrap distribution. From the sorted list, we see the lower end of the bootstrap confidence interval will be 212.94. Moving to the upper end, we find the 78th observation in this distribution is 274.95. Our bootstrap confidence interval for the mean 2002 Seattle house price is then \$212,940 to \$274,950. This is slightly narrower than the interval found in Example 14.4, which was \$214,665 to \$284,439.







The BOOTCORR Program

```
Program:BOOTCORR
:dim(L1)→N
:ClrList L3,L4,L5
:Disp "1=CORRELATION","2=REG. SLOPE"
:Input W
:Disp "NO. OF RESAMPLES"
:Input B
:Disp "CONF. LEVEL"
:Input R
:For(I,1,B)
:randInt(1,N,N)→L6
:For(J,1,N)
:L1(L6(J))→L4(J)
:L2(L6(J))→L5(J)
:End
:LinReg(ax+b) L4,L5
:If W=1
:Then
:r→L3(I)
:Else
:a→L3(I)
:End
:End
:LinReg(ax+b) L1,L2
:If W=1
:Then
:r→X
:Else
:a→X
:End
:ClrList L4,L5,L6
:1-Var Stats L3
:"tcdf(0,X,N-1)"→Y1
:solve(Y1-R/2,X,2)→Q
:ClrHome
:If W=1
:Then
:Disp "SAMPLE CORR."
:Else
:Disp "REG. SLOPE"
:End
:Disp X
:Disp "BOOT SE"
:Disp Sx
:Output(6,2,"CONF. INTERVAL")
:If W=1
:Then
:Output(7,2,max(−1,X-Q*Sx))
:Output(8,2,min(X+Q*Sx,1))
:Else
:Output(7,2,X-Q*Sx)
:Output(8,2,X+Q*Sx)
:End
```

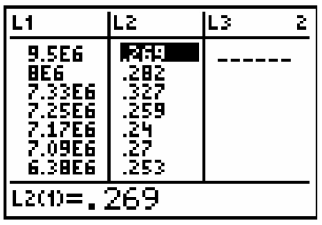
The **BOOTCORR** program can be used to bootstrap the correlation coefficient or the regression slope for paired sample data. Before executing the program, enter the random sample data into lists L1 and L2. When prompted, enter **1** to bootstrap the correlation coefficient or enter **2** to bootstrap the regression slope. Then enter the desired number of resamples and the confidence level. The resampled statistics are stored in list L3, and the statistic of the original sample data is displayed along with the bootstrap standard error and the confidence interval.

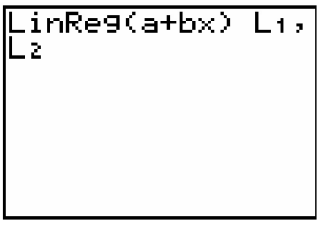
Example 14.8 Baseball Salaries and Batting Averages. The following table gives the 2002 salaries and career batting averages for 50 randomly selected MLB players (excluding pitchers).

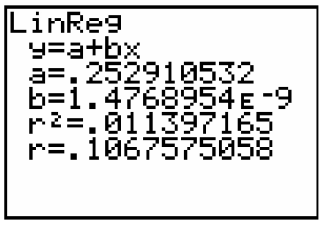
Player	Salary	Avg	Player	Salary	Avg	Player	Salary	Avg
1.	$9,500,000	0.269	18.	$3,450,000	0.242	35.	$630,000	0.324
2.	$8,000,000	0.282	19.	$3,150,000	0.273	36.	$600,000	0.200
3.	$7,333,333	0.327	20.	$3,000,000	0.250	37.	$500,000	0.214
4.	$7,250,000	0.259	21.	$2,500,000	0.208	38.	$325,000	0.262
5.	$7,166,667	0.240	22.	$2,400,000	0.306	39.	$320,000	0.207
6.	$7,086,668	0.270	23.	$2,250,000	0.235	40.	$305,000	0.233
7.	$6,375,000	0.253	24.	$2,125,000	0.277	41.	$285,000	0.259
8.	$6,250,000	0.238	25.	$2,100,000	0.227	42.	$232,500	0.250
9.	$6,200,000	0.300	26.	$1,800,000	0.307	43.	$227,500	0.278
10.	$6,000,000	0.247	27.	$1,500,000	0.276	44.	$221,000	0.237
11.	$5,825,000	0.213	28.	$1,087,500	0.216	45.	$220,650	0.235
12.	$5,625,000	0.238	29.	$1,000,000	0.289	46.	$220,000	0.243
13.	$5,000,000	0.245	30.	$950,000	0.237	47.	$217,500	0.297
14.	$4,900,000	0.276	31.	$800,000	0.202	48.	$202,000	0.333
15.	$4,500,000	0.268	32.	$750,000	0.344	49.	$202,000	0.301
16.	$4,000,000	0.221	33.	$720,000	0.185	50.	$200,000	0.224
17.	$3,625,000	0.301	34.	$675,000	0.234			

(a) Calculate the sample correlation between salary and average.
(b) Bootstrap the correlation and give a 95% confidence interval for the correlation.
(c) Calculate the least-squares regression line to predict batting average from salary. Give the traditional 95% *t* confidence interval for the slope of the regression line.
(d) Bootstrap the regression model. Give a 95% bootstrap *t* confidence interval and a 95% bootstrap percentile confidence interval for the regression slope.

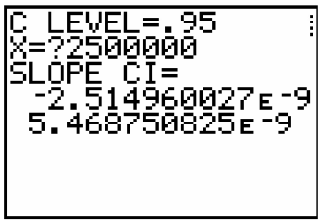
Solution. (a) and (c) We can find the sample correlation, least-squares regression line, and traditional confidence interval for the slope using techniques from Chapter 10. First, we enter the data into lists, say lists L1 and L2. Next, we use LinReg(a+bx) (item 8 from the STAT CALC menu) and enter the command LinReg(a+bx) L1,L2. We obtain a sample correlation of r = 0.1067575. The regression line is given as $BatAvg = .253 + 1.477 * 10^{-9} Salary$. Thus, we see that the slope appears to be 0.



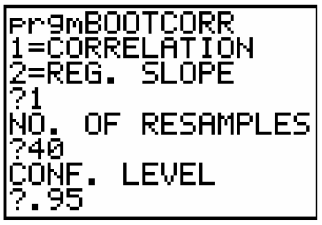


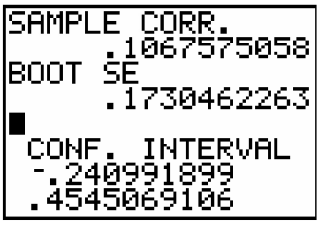


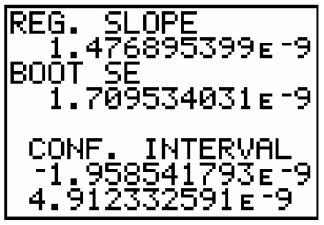
We can find a confidence interval for the slope using the **REGINF** program. After running the program, we find a 95% *t* confidence interval for the slope of the regression line to be $(2.515\times10^{-9}, 5.469\times10^{-9})$.



(b) and (d) We now execute the **BOOTCORR** program twice using 40 resamples each time. In the first running, we enter **1** to bootstrap the correlation. In the second running, we enter **2** to bootstrap the regression slope.

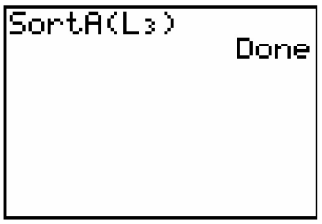


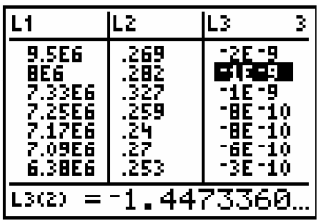


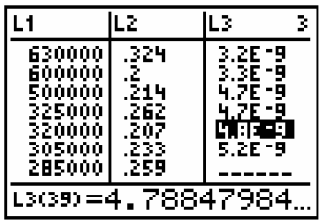


In the first running, we obtain a 95% bootstrap interval for the correlation of (–0.241, 0.456). In the second running, we obtain a 95% bootstrap interval for the regression slope of $(-1.956\times10^{-9}, 4.912\times10^{-9})$.

The 40 resample regression slopes are stored in list `L3` after option 2 of the **BOOTCORR** program ends. So next, we enter the command `SortA(L3)` to resort the list in increasing order. Because .025*40 = 1, the endpoints of the 95% percentile confidence interval are the 2nd and 39th elements of the sorted list. The interval here is $(-1.447\times10^{-9}, 4.788\times10^{-9})$.







14.5 Significance Testing Using Permutation Tests

We conclude this chapter with two programs that will simulate some of the bootstrap permutation tests. The **BOOTTEST** program performs a permutation test for the difference in means, and the **BTPRTEST** program performs a permutation test for either the difference in paired means or for the correlation.

Before executing the **BOOTTEST** program, enter data from the first population into list L1 and enter data from the second population into list L2. When prompted, enter **1**, **2**, or **3** to designate the desired alternative $\mu_1 < \mu_2$, $\mu_1 > \mu_2$, or $\mu_1 \neq \mu_2$. The resampled differences in permuted means are ordered and then stored in list L3. The program displays the difference between the original sample means $\bar{x}_1 - \bar{x}_2$ and the p-value.

The BOOTTEST Program

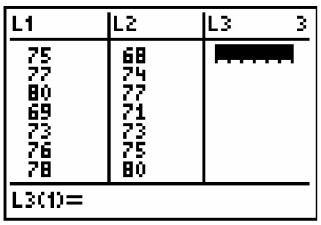
```
Program:BOOTTEST
:Disp "1 = ALT. <","2 = ALT. >","3 = ALT. _"
:Input C
:Disp "NO. OF RESAMPLES"
:Input B
:dim(L1)→N
:dim(L2)→M
:augment(L1,L2)→L1
:ClrList L3
:For(S,1,B)
:ClrList L6
:seq(J,J,1,N+M)→L4
:For(I,1,M)
:ClrList L5
:randInt(1,N+M-I+1)→A
:L4(A)→L6(I)
:1→K
:While K<A
:L4(K)→L5(K)
:1+K→K
:End
:A→K
:While K≤(N+M-I)
:L4(K+1)→L5(K)
:1+K→K
:End
:L5→L4
:End
:sum(seq(L1(L5(J)),J,1,N))/N→X
:sum(seq(L1(L6(J)),J,1,M))/M→Y
:X-Y→L3(S)
:End
:ClrList L4
:seq(L1(I),I,1,N)→L1
:mean(L1)→X
:mean(L2)→Y
:SortA(L3)
:X-Y+1→L3(B+1)
:0→K
:While L3(K+1) <(X-Y)
:K+1→K
:End
:K/B→P
:While L3(K+1) ≤ (X-Y)
:K+1→K
:End
:(B-K)/B→Q
:seq(L3(I),I,1,B)→L3
:ClrHome
:Disp "DIFF IN MEANS"
:Disp X-Y
:Disp "P VALUE"
:If C=1
:Then
:Disp P
:Else
:If C=2
:Then
:Disp Q
:Else
:Disp 2*min(P,Q)
:End
:End
:Stop
```

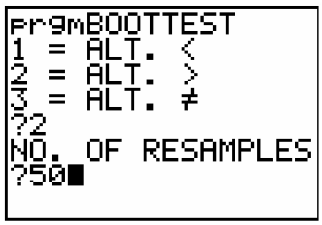
Example 14.9 French Fry Discrimination? A fast-food-restaurant customer complains that people 60 years old or older are given fewer french fries than people under 60. The owner responds by gathering data without knowledge of the employees. Below are the data on the weight of french fries (in grams) from random samples of the two groups of customers. Perform a permutation test using an appropriate alternative hypothesis and give the *p*-value.

Age < 60:	75	77	80	69	73	76	78	74	75	81
Age ≥ 60:	68	74	77	71	73	75	80	77	78	72

Solution. We let μ_1 be the average weight of french fries served to customers under age 60, and let μ_2 be the average weight served to customers age 60 or older. We shall test H_0: $\mu_1 = \mu_2$ with the alternative H_A: $\mu_1 > \mu_2$.

First, we enter the sample weights for the under age 60 customers into list L1 and the weights for the other group into list L2. Then we execute the **BOOTTEST** program by entering **2** to designate the alternative $\mu_1 > \mu_2$. Below are the results from 50 resamples.

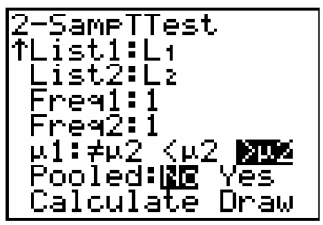


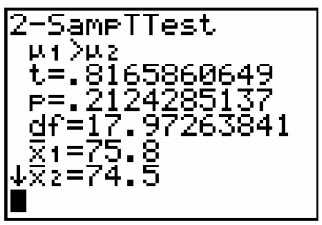




The 50 resampled differences in permuted mean are stored in list L3. The difference in the original sample means is $x_1 - x_2 = 1.3$, and 16% of the means in L3 are greater than 1.3. Based on this sample of permutations, if $\mu_1 = \mu_2$, then there would be about a 16% chance of $\bar{x}_1 - \bar{x}_2$ being as high as 1.3 with samples of these sizes. This *p*-value of 0.16 does not provide enough evidence to reject H_0; it does not appear that the older individuals are getting fewer french fries.

Using the 2-SampTTest feature, we see that the traditional two-sample *t* test gives a *p*-value of 0.21243 and leads us to the same conclusion.





Permutation Test for Paired Data

When we have paired sample data, such as "before and after" measurements, then we often consider the differences in measurements as one sample. The hypothesis test H_0: $\mu_1 = \mu_2$ then becomes H_0: $\mu_1 - \mu_2 = 0$ and may be tested with this one sample using a traditional *t* test or with many resamples using the permutation test. The **BTPRTEST** program may be used for this permutation test. It also may be used to test whether the correlation equals 0.

The BTPRTEST Program

```
Program:BTPRTEST
:Disp "1 = PAIRED MEAN","2 = CORRELATION"
:Input T
:ClrHome
:Disp "1 = ALT. <","2 = ALT. >","3 = ALT. _"
:Input C
:Disp "NO. OF RESAMPLES"
:Input B
:dim(L1)→N
:ClrList L3
:For(I,1,B)
:ClrList L5,L6
:If T=2
:Then
:L1→L5
:randInt(1,N,N) →L4
:For(J,1,N)
:L2(L4(J)) →L6(J)
:End
:LinReg(ax+b) L5,L6
:r→L3(I)
:Else
:For(J,1,N)
:randInt(0,1) →A
:If A=0
:Then
:L1(J) →L5(J)
:L2(J) →L6(J)
:Else
:L1(J) →L6(J)
:L2(J) →L5(J)
:End
:End
:If T=1
:Then
:mean(L5)-mean(L6)→L3(I)
:End
:End
:If T=1
:Then
:mean(L1)-mean(L2)→X
:Else
:LinReg(ax+b) L1,L2
:r→X
:End
:SortA(L3)
:X+1→L3(B+1)
:0→K
:While L3(K+1)<X
:K+1→K
:End
:K/B→P
:While L3(K+1)≤X
:K+1→K
:End
:(B-K)/B→Q
:seq(L3(I),I,1,B) →L3
:ClrHome
:If T=1
:Then
:Disp "DIFF. IN MEANS"
:Else
:Disp "SAMPLE CORR."
:End
:Disp X
:Disp "P VALUE"
:If C=1
:Then
:Disp P
:Else
:If C=2
:Then
:Disp Q
:Else
:Disp 2*min(P,Q)
:End
:End
:Stop
```

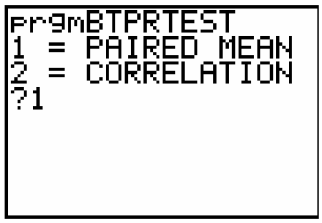
Before executing the **BTPRTEST** program, enter the data into lists L1 and L2. When prompted, enter **1** or **2** to designate the desired test, then enter **1**, **2**, or **3** to designate the desired alternative. The resampled permuted pair differences in mean (or correlation) are ordered and stored in list L3. The statistic from the original paired sample is displayed along with the *p*-value of the permutation test.

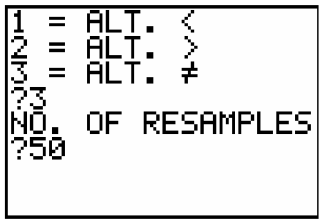
Example 14.10 Are the Technicians Consistent? Below is the data from Example 7.6 (page 85) that gives the total body bone mineral content of eight subjects as measured by two different X-ray machine operators. Perform a matched pairs permutation test to decide whether or not the two operators have the same mean.

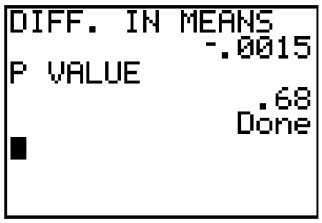
	Subject							
Operator	1	2	3	4	5	6	7	8
1	1.328	1.342	1.075	1.228	0.939	1.004	1.178	1.286
2	1.323	1.322	1.073	1.233	0.934	1.019	1.184	1.304

Solution. We let μ_1 be the average measurement from Operator 1 and let μ_2 be the average measurement from Operator 2. We shall test H_0: $\mu_1 = \mu_2$ with the alternative H_A: $\mu_1 \neq \mu_2$.

First, we enter the measurements into lists L1 and L2. Then we execute the **BTPRTEST** program by entering **1** to designate a paired mean, and then entering **3** for the alternative $\mu_1 \neq \mu_2$. Next, we display the results from 50 resamples.



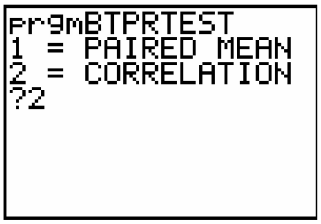


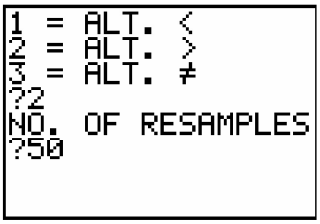


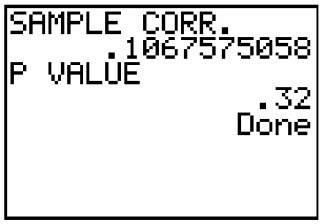
The 50 resampled permuted pair differences in mean are stored in list L3. The difference in the original sample means is $\bar{x}_1 - \bar{x}_2 = -0.0015$ and 68%/2 = 34% of the means in L3 are less than –0.0015. This large *p*-value means that we do not have significant evidence to argue that there is a difference in the operator's mean measurement.

Example 14.11 More on Baseball Salaries. Use the data from Example 14.8 to test whether the correlation between salary and batting average is greater than 0.

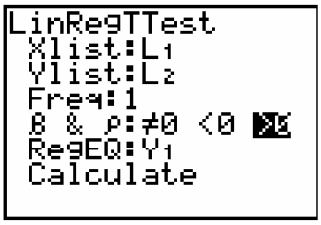
Solution. First, we enter the data into lists L1 and L2. Then to test the hypothesis H_0: $\rho = 0$ with alternative H_A: $\rho > 0$, we execute option 2 of the **BTPRTEST** program using the second alternative with 50 resamples.

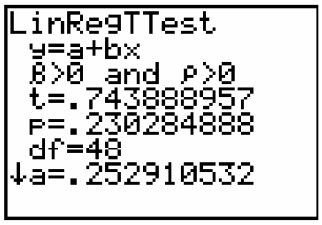






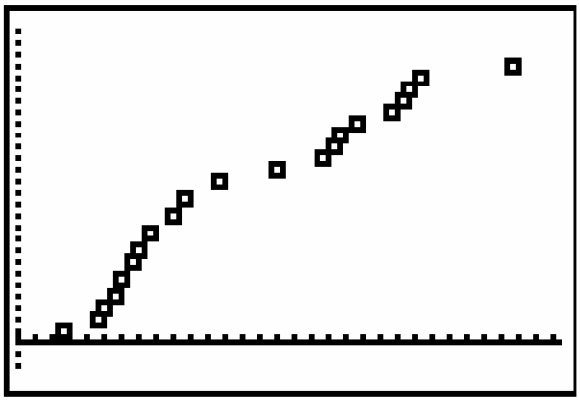
Upon running the program, the 50 resample correlations are stored in list L3. The sample correlation is $r = 0.1067575$. In this case, 16 of the resample correlations were greater than r, which gives us a one-sided p-value of 0.32. Thus, we do not have significant evidence to reject H_0 with this sample based upon 50 resamples. Our p-value of 0.32 compares favorably with the p-value of 0.23 obtained with the traditional linear regression t test.





CHAPTER

15



Nonparametric Tests

15.1	The Wilcoxon Rank Sum Test
15.2	The Wilcoxon Signed Rank Test
15.3	The Kruskal-Wallis Test

Introduction

In this chapter, we provide some supplementary programs for performing several nonparametric hypothesis tests.

These tests relax the assumption of Normal distribution of the data (or sample mean). They are less powerful than parametric tests, since they do not use all the information in the data but only the ranks (sorted order statistics); one has information on the smallest, next smallest, and so forth, but not the size of the differences.

The Rank Sum test is a stand-in for a two sample *t* test; the signed rank for a paired samples test, and the Kruskal-Wallis for one-way ANOVA.

15.1 The Wilcoxon Rank Sum Test

We first provide the program **RANKSUM** to perform the Wilcoxon rank sum test on data from two populations. To execute the program, we must enter the data into lists L1 and L2. The program sorts each list, then merges and sorts the lists into list L3. Then it stores the rank of each measurement in L3 next to it in list L4. All sequences of ties are assigned an average rank.

The Wilcoxon test statistic W is the sum of the ranks from L1. Assuming that the two populations have the same continuous distribution (and no ties occur), then W has a mean and standard deviation given by

$$\mu = \frac{n_1(N+1)}{2} \quad \text{and} \quad \sigma = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

where n_1 is the sample size from L1, n_2 is the sample size from L2, and $N = n_1 + n_2$.

We test the null hypothesis H_0: there is no difference in distributions. A one-sided alternative is H_A: the first population yields higher measurements. We use this alternative if we expect or see that W is a much higher sum than its expected sum of ranks μ. In this case, the p-value is given by a normal approximation. We let $W \sim N(\mu, \sigma)$ and compute the right-tail $P(X \geq W)$ (using the continuity correction if W is an integer).

If we expect or see that W is the much lower sum than its expected sum of ranks μ, then we should use the alternative H_A: first population yields lower measurements. In this case, the p-value is given by the left-tail $P(X \leq W)$, again using continuity correction if needed. If the two sums of ranks are close, then we could use a two-sided alternative H_A: there is a difference in distributions. In this case, the p-value is given by twice the smallest tail value: $2 * P(X \geq W)$ if $W > \mu$, or $2 * P(X \leq W)$ if $W < \mu$.

The **RANKSUM** program displays the expected sum of ranks from the first list and the actual sums of the ranks from L1 and L2. It also displays the smallest tail value created by the test statistic. That is, it displays $P(X \geq W)$ if $W > \mu$ and it displays $P(X \leq W)$ if $W < \mu$; for a two-tail alternate hypothesis, multiply the given p-value by 2. Conclusions for any alternative then can be drawn from this value. We note that if there are ties, then the validity of this test is questionable.

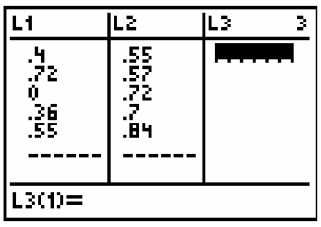
Example 15.1 Retelling Stories. Below are language usage scores of kindergarten students who were classified as high-progress readers or low-progress readers when asked to retell a story that had been read to them. Is there evidence that the scores of high- progress readers are higher than those of low-progress readers? Carry out a two-sample t test. Then carry out the Wilcoxon rank sum test and compare the conclusions for each test.

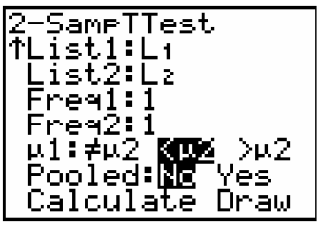
Child	Progress	Score
1	high	0.55
2	high	0.57
3	high	0.72
4	high	0.70
5	high	0.84
6	low	0.40
7	low	0.72
8	low	0.00
9	low	0.36
10	low	0.55

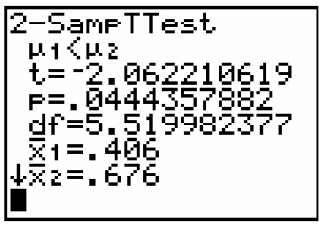
The RANKSUM Program

```
PROGRAM:RANKSUM
:ClrList L4
:SortA(L1):SortA(L2)
:augment(L1,L2) STO▶L3
:SortA(L3)
:dim(L1) STO▶M:dim(L2) STO▶N
:L3 STO▶L6
:L3(1)-1 STO▶L3(M+N+1)
:1 STO▶B
:1STO▶I
:Lbl 1
:While I<(M+N)
:If L3(I)<L3(I+1)
:Then
:BSTO▶L4(I)
:1+ISTO▶I
:1+BSTO▶B
:Goto 1
:Else
:1STO▶J
:BSTO▶S
:Lbl 2
:While L3(I)=L3(I+J)
:S+B+JSTO▶S
:1+JSTO▶J
:Goto 2
:End
:S/JSTO▶T
:For(K,0,J-1)
:TSTO▶L4(I+K)
:End
:I+JSTO▶I:B+JSTO▶B
:Goto 1
:End
:End
:If I=M+N
:Then
:M+NSTO▶L4(I)
:End
:1STO▶I:0STO▶S:0STO▶J
:Lbl 3
:While I≤M
:Lbl 4
:If L1(I)=L3(I+J)
:Then
:S+L4(I+J) STO▶S
:Else
:1+JSTO▶J
:Goto 4
:End
:I+1STO▶I
:Goto 3
:End
:L6STO▶L3
:ClrList L6
:(M+N)(M+N+1)/2STO▶R
:M*(M+N+1)/2STO▶U
:√(M*N*(M+N+1)/12) STO▶D
:(abs(S-U)-.5)/DSTO▶Z
:If int(S)≠S:(abs(S-U))/DSTO▶Z:End
:0.50-normalcdf(0,Z,0,1) STO▶P
:Disp "EXPECTED 1ST SUM"
:Disp .5*M*(M+N+1)
:Disp "SUMS OF RANKS"
:Disp {S,R-S}
:If S=U:Disp "NO DIFFERENCE"
:If S<U:Disp "LEFT TAIL",round(P,4)
:If S>U:Disp "RIGHT TAIL",round(P,4)
:Stop
```

Solution. First, we enter the five low-progress scores into L1 and the five high-progress scores into L2. Then we use the 2-SampTTest from the STAT TESTS menu to test the hypothesis H_0: $\mu_1 = \mu_2$ versus the alternative H_A: $\mu_1 < \mu_2$.

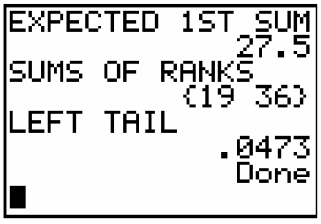


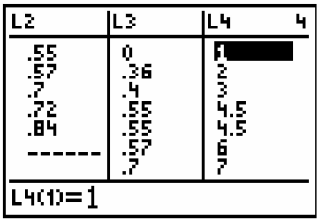


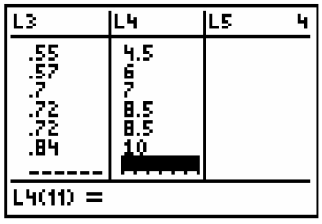


We obtain a t statistic of –2.06221 and a p-value of 0.0444 for the one-sided alternative. With the rather small p-value, we have significant evidence to reject H_0 and say that the average score of all high-progress readers is higher than the average score of all low-progress readers. For if H_0 were true, then there would be only a 0.0444 probability of obtaining a high-progress sample mean that is so much larger than the low-progress sample mean (0.676 compared to 0.406).

Now for the Wilcoxon rank sum test, we use H_0: the distribution is the same for both groups versus H_A: high-progress readers score higher when retelling the story. The Wilcoxon test statistic is the sum of ranks from L1 in which we entered the low-progress scores. With the data entered into lists L1 and L2, we now execute the **RANKSUM** program, and then observe the sorted data and ranks in lists L3 and L4.







The sum of the ranks from the low-progress readers is 19, which is lower than the expected average of μ = (5*11)/2 = 27.5. According to the Wilcoxon test, if the distributions were the same, then there would be only be a 0.0473 probability (from the left-tail value) of the low- progress sum of ranks being so much smaller than the expected average of 27.5. Therefore, we should reject H_0 in favor of the alternative; high-progress readers are better at retelling stories told to them.

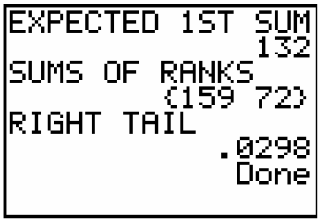
In this case, the Wilcoxon p-value is slightly higher than the t test p-value; however, both are low enough to result in the same conclusion.

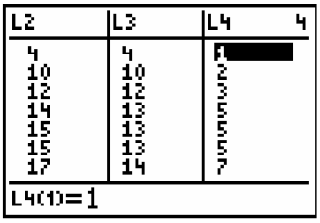
Example 15.2 Logging in the Rain Forest. Below is a comparison of the number of tree species in unlogged plots in the rain forest of Borneo with the number of species in plots logged eight years earlier.

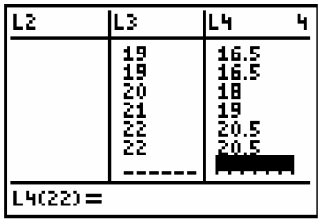
Unlogged	22	18	22	20	15	21	13	13	19	13	19	15
Logged	17	4	18	14	18	15	15	10	12			

Does logging significantly reduce the mean (median) number of species in a plot after eight years? State the hypotheses, do a Wilcoxon rank sum test, and state your conclusion.

Solution. We will test the hypothesis H_0: there is no difference in medians (or distributions) versus the alternative H_A: the unlogged median is higher. To do so, we first enter the unlogged measurements into list L1 and enter the logged measurements into list L2. Then we execute the **RANKSUM** program which produces the following results:







We note that there are 21 total measurements with 12 unlogged measurements. If there were no difference in distributions or medians, then we would expect the sum of ranks from L1 to be $\mu = 12*22/2 = 132$. But if there were no difference in medians, then there would be only a 2.98% chance of the sum of ranks from L1 being as high as 159. This low *p*-value gives significant evidence to reject H_0 in favor of the alternative; there are more species of trees in the unlogged plots.

15.2 The Wilcoxon Signed Rank Test

Here we provide the **SIGNRANK** program to perform the Wilcoxon signed rank test on data sets of size *n* from two populations (this can also be used to test a single median). To execute the program, we must enter the data into lists L1 and L2. The program will sort the absolute value of the differences L2 – L1 into list L3, but it will disregard any zero differences. The sample size *n* is decreased so as to count only the nonzero differences. Then the program puts the rank of each measurement in L3 next to it in L4. All sequences of ties are assigned an average rank.

The Wilcoxon test statistic *W* is the sum of the ranks from the positive differences. Assuming that the two populations have the same continuous distribution (and no ties occur), then *W* has a mean and standard deviation given by

$$\mu = \frac{n(n+1)}{4} \quad \text{and} \quad \sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

The SIGNRANK Program

```
PROGRAM:SIGNRANK
:ClrList L3,L4,L5
:0→L3(1):0→L4(1):0→L5(1)
:0→S:0→R:0→U:0→V
:1→J
:For(I,1,dim(L1),1)
:If L2(I)-L1(I)≠0
:Then
:abs(L2(I)-L1(I))→L3(J)
:1+J→J
:End:End
:If L3(1)>0
:Then
:SortA(L3):dim(L3)→N
:L3→L6:L3(1)-1→L3(N+1)
:1→B:1→I
:Lbl 1
:If I<N
:Then
:If L3(I)<L3(I+1)
:Then
:B→L4(I)
:1+I→I:1+B→B
:Goto 1
:Else
:1→J:B→S
:Lbl 2
:If L3(I)=L3(I+J)
:Then
:S+B+J→S:1+J→J
:Goto 2
:End
:S/J→T
:For(K,0,J-1)
:T→L4(I+K)
:End
:I+J→I:B+J→B
:End
:Goto 1
:End
:If I=N
:Then
:N→L4(I)
:End
:0→J
:For(I,1,dim(L1))
:If (L2(I)-L1(I))>0
:Then
:1+J→J:abs((L2(I)-L`(I))→L5(J)
:End:End
:SortA(L5)
:1→I:0→J
:If L5(1)>0
:Then
:Lbl 3
:While I≤dim(L5)
:Lbl 4
:If L5(I)=L3(I+J)
:Then
:V+L4(I+J)→V
:Else
:1+J→J
:Goto 4
:End
:I+1→I
:Goto 3
:End:End
:L6→L3:ClrList L6
:N(N+1)/2→R:N(N+1)/4→U
:√(N(N+1)(2N+1)/24)→D
:(abs(V-U)-.5)/D→Z
:If int(V)≠V:(abs(V-U))/D→Z:End
:0.50-normalcdf(0,Z,0,1)→P
:Disp "EXP. + SUM"
:Disp N*(N+1)/4
:Disp "SUMS -,+ RANKS"
:Disp {R-V,V}
:If V=U:Disp "NO DIFFERENCE"
:If V<U:Disp "LEFT TAIL",round(P,4)
:If V>U:Disp "RIGHT TAIL",round(P,4)
:Stop
```

We test the null hypothesis H_0: there is no difference in distributions. A one-sided alternative may be H_A: the second population yields higher measurements. We use this alternative if we expect or see that W is a much higher sum, which means that there were more positive differences in L2 – L1. In this case, the p-value is given by a normal approximation. We let $X \sim N(\mu, \sigma)$ and compute the right-tail $P(X \geq W)$ (using a continuity correction if W is an integer).

If we expect or see that W is the much lower sum, then there were more negative differences. Now we should use the alternative H_A: the second population yields lower measurements. In this case, the p-value is given by the left-tail $P(X \leq W)$, again using continuity correction if needed. If the two sums of ranks are close, we could use a two-sided alternative H_A: there is a difference in distributions. In this case, the p-value is given by twice the smallest tail value.

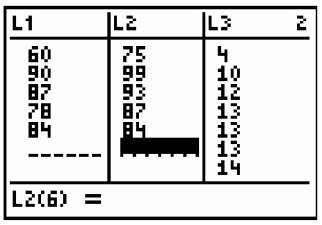
The **SIGNRANK** program displays the sums of the ranks of the negative differences and of the positive differences as well as the smallest tail value created by the test statistic. That is, it displays $P(X \geq W)$ if $W > \mu$, or $P(X \leq W)$ if $W < \mu$. Conclusions for any alternative can then be drawn from this value. Again, we note that if there were ties, then the validity of this test is questionable.

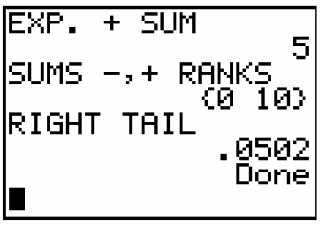
Example 15.3 Stepping and Heart Rates. A student project asked subjects to step up and down for three minutes. There were two treatments: stepping at a low rate (14 steps per minute) and a medium rate (21 steps per minute). Here are data for heart rates for five subjects and the two treatments. For each subject we have the initial resting heart rate and the heart rate at the end of the exercise. Does exercise at the low rate raise the heart rate significantly? State hypotheses in terms of the median increase in heart rate and apply the Wilcoxon signed rank test.

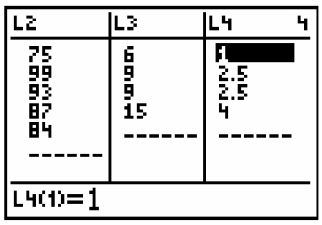
	Low rate		Medium rate	
Subject	**Resting**	**Final**	**Resting**	**Final**
1	60	75	63	84
2	90	99	69	93
3	87	93	81	96
4	78	87	75	90
5	84	84	90	108

Solution. We will test the hypothesis H_0: For the low stepping rate, resting and final heart rates have the same median versus H_A: final heart rates are higher.

Enter the five low rate resting heart rates into list L1 and the five low rate final heart rates into list L2. The alternative means that there should be more positive differences, so that the sum of the positive ranks should be higher. Therefore, the p-value comes from the right-tail probability created by the test statistic. After the data is entered, execute the **SIGNRANK** program.







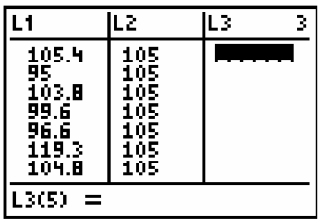
List L3 now contains the ordered absolute values of the four nonzero differences. Their corresponding (averaged) ranks are adjacent in list L4. List L5 contains only the positive differences, which in this case are all four of the differences.

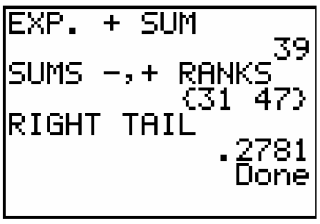
We see that the sum of the ranks of the positive differences is much higher than that of the negative differences. If the medians for each rate were the same, then there would be only a 0.0502 probability of the sum of positive ranks being as high as 10. They were expected to be 5 with the four subjects for which there is a difference. The relatively low *p*-value provides some evidence to reject H_0 and conclude that the median final heart rate is higher for the low rate test.

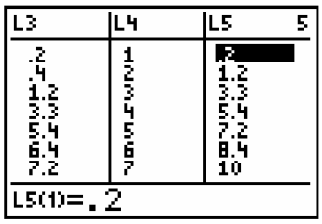
Example 15.4 Radon Detector Accuracy. Below are the readings from 12 home radon detectors exposed to 105 pCi/l of radon. We want to know if these detectors are accurate. Apply the Wilcoxon signed rank test to determine if the median reading from all such home radon detectors differs significantly from 105.

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

Solution. We will test the null hypothesis H_0: median = 105 versus H_A: median ≠ 105. First, we enter the given data into list L1 and then enter 105 into list L2 twelve times. If H_0 were true, then we would expect the sum of ranked positive differences of L2 – L1 to be (12*13) /4 = 39. But H_A implies that this sum of ranked positive differences will be either much higher than 39 or much lower than 39. To test the hypotheses, we execute the **SIGNRANK** program after entering these data into the lists.







Lists L3 and L4 will show that there were 12 nonzero differences in L2 – L1, and list L5 will show that eight of these were positive differences, meaning that there were eight times in which the home radon detector measured below 105.

The right-tail value is given as 0.2781; thus, the *p*-value for the two-sided alternative is 2*0.2781 = 0.5562. If the median home radon measurement were 105, then there would be a 0.5562 probability of the sum of positive ranks being as far away (in either direction) from the expected sum of 39 as the resulting sum of 47 is. Thus, we do not have significant evidence to reject H_0 and can conclude that, "on average," these detectors are accurate.

15.3 The Kruskal-Wallis Test

Our next program, **KRUSKAL**, is for the Kruskal-Wallis test, which simultaneously compares the distribution of more than two populations. We test the null hypothesis H_0: all populations have the same distribution versus the alternative H_A: measurements are systematically higher in some populations. To apply the test, we draw independent SRSs of sizes $n_1, n_2, \ldots, n_I$ from I populations. There are N observations in all. We rank all N observations and let R_i be the sum of the ranks for the ith sample. The Kruskal-Wallis statistic is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} - 3(N+1)$$

When the sample sizes are large and all I populations have the same continuous distribution, then H has an approximate chi-square distribution with $I-1$ degrees of freedom. When H is large, creating a small right-tail *p*-value, then we can reject the hypothesis that all populations have the same distribution.

The KRUSKAL Program

```
PROGRAM:KRUSKAL
:ClrList L3,L4,L5
:dim([B])→L1
:sum(seq([B](1,J),J,1,L1(2)))→L
:1→K
:For(J,1,L1(2))
:For(I,1,[B](1,J))
:[A](I,J)→L3(K)
:1+K→K
:End:End
:SortA(L3)
:L3→L6:L3(1)-1→L3(L+1)
:1→B:1→I
:Lbl 1
:While I<(L)
:B→L4(I)
:1+I→I:1+B→B
:Goto 1
:Else
:1→J:B→S
:Lbl 2
:While L3(I)–L3(I+J)
:S+B+J→S:1+J→J
:Goto 2
:End
:S/J→T
:For(K,0,J-1)
:T→L4(I+K)
:End
:I+J→I:B+J→B
```

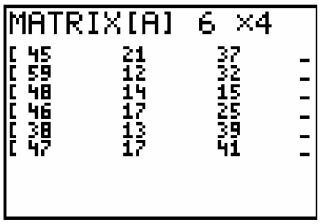
```
:If L3(I)<L3(I+1)
:Then
:If I=L
:Then
:L▶L4(I)
:End
:1▶K
:Lbl 5
:While K≤L1(2)
:ClrList L2
:For(I,1,[B](1,K))
:[A](I,K) ▶L2(I)
:SortA(L2)
:End
:1▶I:0▶S:0▶J
:Lbl 3
:While I≤[B](1,K)
:Lbl 4
:If L2(I)=L3(I+J)
:Goto 1
:End:End
:Then
:S+L4(I+J) ▶S:1+I▶I
:Else
:1+J▶J
:End
:Goto 3
:End
:S▶L5(K):1+K▶K
:Goto 5
:End
:L6▶L3:ClrList L2,L6
:12/L/(L+1)*sum(seq(L5(I)²/
 [B](1,I),I,1,L1(2)))-3(L+1) ▶W
:1-▶χ2cdf(0,W,L/(2)-1) ▶P
:Disp "TEST STAT",W
:Disp "P-VALUE",P
```

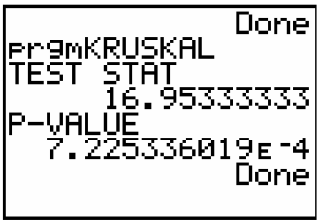
Before executing the **KRUSKAL** program, use MATRX EDIT to enter the data as columns into matrix [A] and to enter the sample sizes as a row into matrix [B].

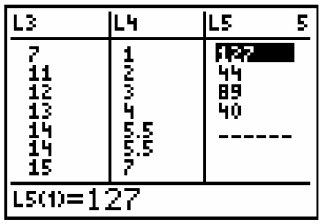
Example 15.5 Are Insects Colorblind? An experiment was conducted to determine if insects were equally attracted by different colors. Sticky boards were placed in a field of oats, and the number of cereal leaf beetles trapped was counted. Use the Kruskal-Wallis test to see if there are significant differences in the numbers of insects trapped by the different board colors.

Board color	Insects trapped					
Lemon yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	37	32	15	25	39	41
Blue	16	11	20	21	14	7

Solution. To execute the **KRUSKAL** program, we must use MATRX EDIT to enter the data into matrix [A] and enter the sample sizes into matrix [B]. First, enter the data into the columns of the 6×4 matrix [A] as you would normally enter data into lists. Next, enter the sample sizes into a 1×4 matrix [B]. Then execute the **KRUSKAL** program.



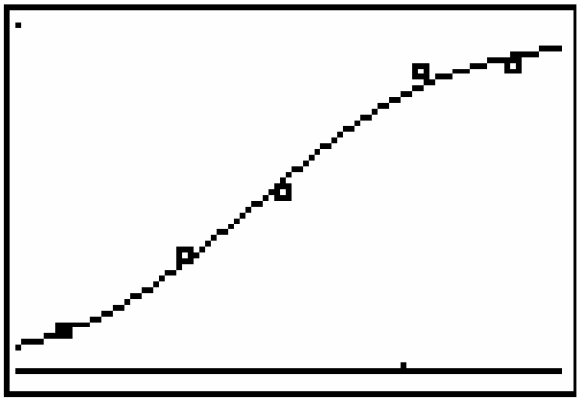




After the program completes, view the entries in lists L3, L4, and L5. List L3 contains the merged, sorted measurements, and list L4 contains their (averaged) ranks. List L5 contains the sum of ranks from each type of color. The low p-value of 0.00072 gives good evidence to reject the hypothesis that all colors yield the same distribution of insects trapped.

CHAPTER

16



Logistic Regression

16.1	The Logistic Regression Model
16.2	Inference for Logistic Regression

Introduction

In this chapter, we give a brief discussion of two types of logistic regression fits. The first type is a linear fit for the logarithm of the odds ratio of two population proportions. The second type is the general logistic fit for several population proportions.

We also discuss the logistic regression model that is commonly used in modeling population sizes.

16.1 The Logistic Regression Model

First, we provide a supplementary program that computes appropriate mathematical odds for a given probability p of an event A. If $p \le 0.50$, then the odds *against* A are given as the ratio $(1-p):p$. If $p > 0.50$, then the odds *in favor of* A are given as the ratio $p:(1-p)$. The probability can be entered either as a decimal or as a fraction. If p is entered as a decimal, then the odds are computed after rounding p to four decimal places; thus some accuracy may be lost.

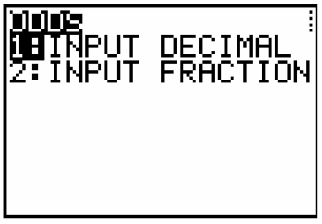
The ODDS Program

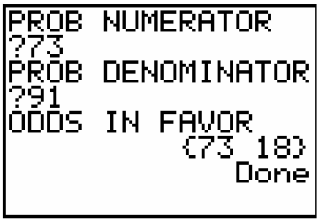
```
PROGRAM:ODDS
:Menu("ODDS","INPUT DECIMAL",
 1,"INPUT FRACTION",2)
:Lbl 1
:Disp "PROBABILITY"
:Input P
:round(P,4)STO▶P
:10000*PSTO▶A
:10000*(1-P) STO▶B
:A/gcd(A,B) STO▶N
:B/gcd(A,B) STO▶D
:Goto 3
:End
:Lbl 2
:Disp "PROB NUMERATOR"
:Input A
:Disp "PROB DENOMINATOR"
:Input B
:A/BSTO▶P
:A/gcd(A,B-A) STO▶N
:(B-A)/gcd(A,B-A) STO▶D
:Lbl 3
:If P>.50
:Then
:Disp "ODDS IN FAVOR"
:Disp {N,D}
:Else
:Disp "ODDS AGAINST"
:Disp {D,N}
:Stop
```

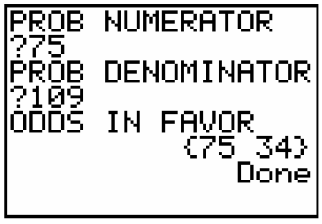
Example 16.1 Employee Stock Options. In a study of 91 high-tech companies and 109 non-high-tech companies, 73 of the high-tech companies and 75 of the non-high-tech companies offered incentive stock options to key employees.

(a) What proportion of high-tech companies offer stock options to their key employees? What are the odds?
(b) What proportion of non-high-tech companies offer stock options to their key employees? What are the odds?
(c) Find the odds ratio using the odds for the high-tech companies in the numerator.

Solution. (a) The proportion of high-tech companies that offer stock options is simply 73/91 = 0.8022. To compute the odds, we can use the **ODDS** program. After bringing up the program, enter the numerator value of **73** followed by the denominator value of **91**. The odds in favor are displayed as 73 to 18.

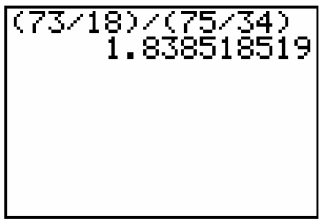






(b) For the non-high-tech companies, the proportion is 75/109 = 0.688. Because 34 of these companies do not offer stock options, the odds are 75 : 34 that such a company does offer stock options to their key employees. This result can be verified with the **ODDS** program.

(c) The odds-in-favor ratio can be computed by simple division of the odds (73/18)/(75/34). Thus, the odds in favor of a high-tech company offering stock options are about 1.8 times more than the odds for a non-high-tech company.



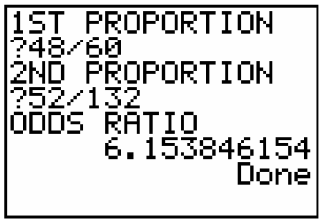
The logistic regression model is always based on the odds in *favor* of an event. It provides another method of studying the odds in favor ratio between two populations. As a lead-in, we now provide a program that specifically computes the odds-in-favor ratio. This program merely automates the division done just above.

The ODDS2 Program

PROGRAM:ODDS2	:Input R
:Disp "1ST PROPORTION"	:Disp "ODDS RATIO"
:Input P	:Disp P/(1-P)/(R/(1-R))
:Disp "2ND PROPORTION"	:Stop

Example 16.2 Gender Bias. In a study on gender bias in textbooks, 48 out of 60 female references were "girl." Also, 52 out of 132 male references were "boy." These two types of references were denoted as juvenile references. Compute the odds ratio for comparing the female juvenile references to the male juvenile references.

Solution. We simply enter the data into the **ODDS2** program that computes the ratio of odds. We see that the odds in favor of a juvenile female response are more than six times the odds in favor of a juvenile male response.



Model for Logistic Regression

The logistic regression model is given by the equation

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where ln is the natural (base e) logarithm and x is either 1 or 0 to designate the explanatory variable. We now provide a program that computes and displays the regression coefficients for the fit $\ln(\text{ODDS}) = b_0 + b_1 x$ as well as the odds ratio e^{b1}.

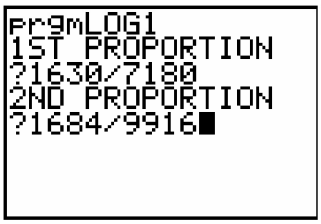
The LOG1 Program

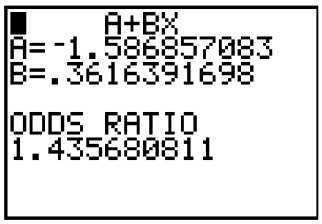
```
PROGRAM:LOG1
:Disp "1ST PROPORTION"
:Input P
:Disp "2ND PROPORTION"
:Input R
:ln(R/(1-R))STO▶A
:ln(P/(1-P))-ASTO▶B
:ClrHome
:Output(1,6,"A+BX")
:Output(2,1,"A=")
:Output(2,3,A)
:Output(3,1,"B=")
:Output(3,3,B)
:Output(5,1,"ODDS RATIO")
:Output(6,1,e^(B))
:Stop
```

Example 16.3 Binge Drinking on Campus. The table below gives data on the numbers of men and women who responded "Yes" to being frequent binge drinkers in a survey of college students. Find the coefficients for the logistic regression model and the odds ratio of men to women.

Population	X	n
Men	1630	7180
Women	1684	9916

Solution. We simply enter the data into the **LOG1** program to obtain the equation $\log(ODDS) = -1.59 + .36x$. The odds ratio is also displayed.





16.2 Inference for Logistic Regression

To compute confidence intervals for the slope β_1 and the odds ratio of the logistic regression model, we can use the **ODDSINT** program that follows. To execute the program, separately enter the numerators and denominators of the two sample proportions and the desired level of confidence.

The ODDSINT Program

```
Program:ODDSINT
:Disp "1ST NO. OF YES"
:Input M
:Disp "1ST SAMPLE SIZE"
:Input P
:Disp "2ND NO. OF YES"
:Input N
:Disp "2ND SAMPLE SIZE"
:Input Q
:Disp "CONF. LEVEL"
:Input L
:√(1/M+1/(P-M)+1/N+1/(Q-N))STO▶S
:invNorm((L+1)/2,0,1) STO▶R
:ln(M/P/(1-M/P))-ln(N/Q/(1-N/Q)) STO▶B
:ClrHome
:Disp "STD. ERROR"
:Disp S
:Disp "SLOPE INTERVAL"
:Disp {round(B-R*S,4),round(B+R*S,4)}
:Disp "ODDS RATIO INT."
:Disp {round(e^(B-R*S),4),
 round(e^(B+R*S), 4)}
:Stop
```

To test the hypothesis that an odds ratio equals 1, we equivalently can test whether the logistic regression model coefficient b_1 equals 0. To do so, we use the p-value given by

$$P(\chi 2(1) \geq (b_1 / SE(b_1)^2))$$

where $SE(b_1)$ is the standard error of the coefficient b_1. The **ODDSTEST** program that follows computes this p-value upon entering the values of the two proportions under consideration and the value of the standard error $SE(b_1)$.

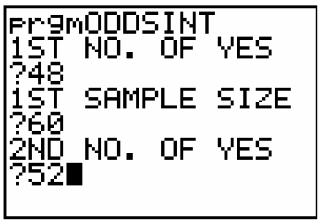
The ODDSTEST Program

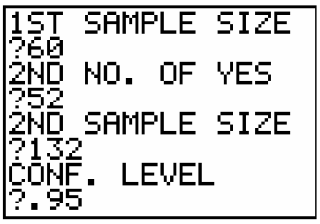
```
Program:ODDSTEST
:Disp "1ST PROPORTION"
:Input P
:Disp "2ND PROPORTION"
:Input R
:Disp "ST.ERROR OF B1"
:Input S
:ln(R/(1-R))STO▶A
:ln(P/(1-P))-ASTO▶B
:(B/S)^2STO▶Z
:1-χ2cdf(0,Z,1) STO▶P
:ClrHome
:Disp "TEST STAT",Z
:Disp "P-VALUE",P
```

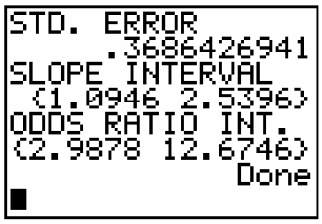
Example 16.4 More Gender Bias. In the study on gender bias in textbooks from Example 16.2, 48 out of 60 female references were "girl" and 52 out of 132 male references were "boy."

(a) Give a 95% confidence interval for the slope.
(b) Calculate the $\chi 2$ statistic for testing the null hypothesis that the slope is zero and give the approximate p-value.

Solution. (a) We execute the **ODDSINT** program and input the proportion information.

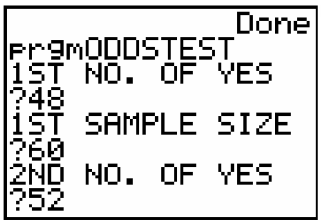


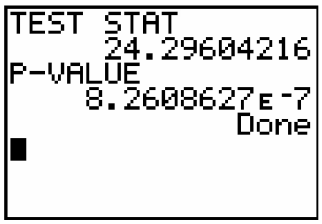




We obtain a 95% confidence interval for the slope of (1.0946, 2.5396) which is equivalent to an odds ration interval of (2.9878, 12.6746). Because 0 is not in the slope interval and 1 is not in the odds ratio interval, we have some evidence to reject the null hypothesis that the slope is zero; the odds of juvenile references are not equal between the two genders.

(b) We execute the **ODDSTEST** program and input the information from the samples.





The $\chi 2$ statistic of 24.296 yields a very low p-value of about 8.26×10^{-7} that gives significant evidence to reject the null hypothesis that the slope equals zero (or that the odds ratio is 1).

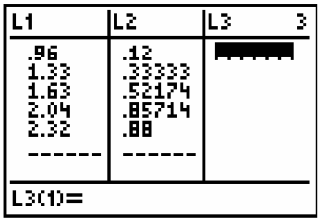
The Logistic Curve

A general logistic curve is given by the function $p = \dfrac{c}{1+ae^{-bx}}$. Such a fit can be obtained with the **Logistic** command from the STAT CALC menu. Following is an example to illustrate this fit.

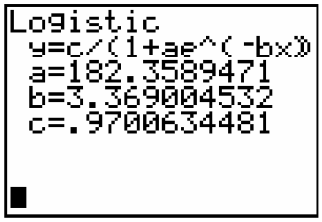
Example 16.5 Insecticide Effectiveness. An experiment was designed to examine how well an insecticide kills a certain type of insect. Find the logistic regression curve for the proportion of insects killed as a function of the insecticide concentration.

Concentration	Number of insects	Insects killed
0.96	50	6
1.33	48	16
1.63	46	24
2.04	49	42
2.32	50	44

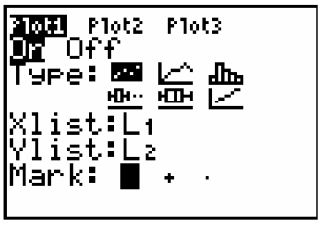
Solution. First, enter the data into the STAT EDIT screen with the concentrations in list L1 and the proportions of insects killed in list L2. Next, enter the command Logistic L1,L2,Y1 to compute the regression fit and to store the equation in function Y1. If you have forgotten, to get to Y1, press [VARS], [▶] to Y-VARS, [ENTER] for Function, [ENTER] for Y1.

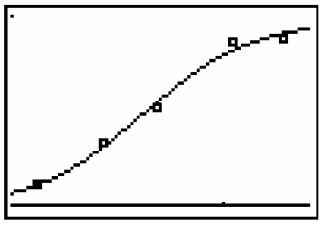






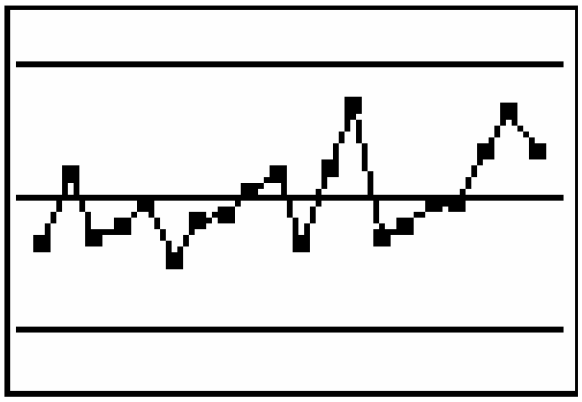
We obtain a logistic regression fit of $y = \dfrac{.97}{1+182.359e^{-3.369x}}$. If desired, we can make a scatterplot of the data along with the logistics regression curve. This clearly shows the increase in effectiveness of the insecticide until, finally, increasing the concentration does little to increase the effectiveness.





CHAPTER

17



Statistics for Quality: Control and Capability

17.1	Statistical Process Control
17.2	Using Control Charts
17.3	Process Capability Indexes
17.4	Control Charts for Sample Proportions

Introduction

In this chapter, we provide several programs for computing control limits, graphing control charts, and for computing the capability indices of a process.

17.1 Statistical Process Control

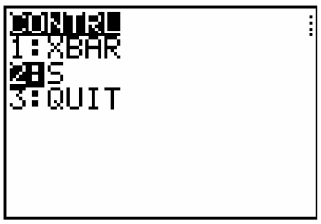
In this section, we provide a program that computes the upper and lower control limits and graphs the control charts for $\bar{x}$ and s.

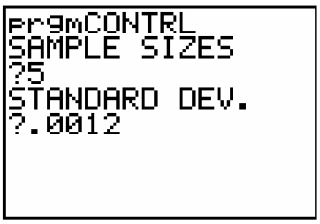
The CONTRL Program

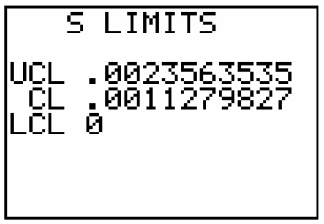
```
PROGRAM:CONTRL
:Menu("CONTRL","XBAR",1,"S",2,"
 QUIT",3)
:Lbl 1
:Disp "SAMPLE SIZES"
:Input N
:Disp "MEAN"
:Input M
:Disp "STANDARD DEV."
:Input S
:"M"STO▶Y1
:"M+3*S/√(N)" STO▶Y2
:"M-3*S/√ (N)" STO▶Y3
:If dim(L1)>0
:Then
:seq(I,I,1,dim(L1))STO▶L3
:0STO▶Xmin
:dim(L1)+1STO▶Xmax
:Xmax STO▶Xscl
:min(min(L1),M-4*S/√ (N)) STO▶Ymin
:max(max(L/),M+4*S/√ (N)) STO▶Ymax
:Ymax-YminSTO▶Yscl
:Plot1(xyLine,L3,L1,▫)
:End
:ClrHome
:Output(1,4,"XBAR LIMITS")
:Goto 4
:Lbl 2
:Disp "SAMPLE SIZES"
:Input N
:Disp "STANDARD DEV."
:Input S
:If int(N/2)=N/2
:Then
:((N/2-1)!)^2*2^(N-2)/(N-2)!*
 √ (2/Π/(N-1)) STO▶C
:Else
: √ (2Π/(N-1))*(N-1)!/((N-1)/2)!
 /((N-3)/2)!/2^(N-1) STO▶C
:End
:"C*S"STO▶Y1
:"C*S+3S√ (1-C^2)"STO▶Y2
:"max(C*S-3S√ (1-C^2),0)" STO▶Y3
:If dim(L2)>0
:Then
:seq(I,I,1,dim(L2))STO▶L3
:0STO▶Xmin
:dim(L2)+1STO▶Xmax
:Xmax STO▶Xscl
:min(min(L2),C*S-4S√ (1-C^2))
 STO▶Ymin
:max(max(L2),C*S+4S√ (1-C^2))
 STO▶Ymax
:Ymax-YminSTO▶Yscl
:Plot1(xyLine,L3,L2, ▫)
:End
:ClrHome
:Output(1,4,"S LIMITS")
:Lbl 4
:PlotsOff
:PlotsOn 1
:AxesOff
:Output(3,1,"UCL")
:Output(3,5,Y2)
:Output(4,1," CL")
:Output(4,5,Y1)
:Output(5,1,"LCL")
:Output(5,5,Y3)
:Lbl 3
:Stop
```

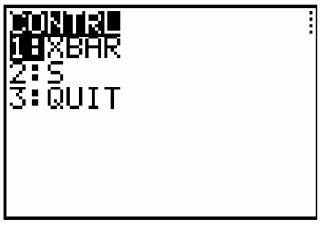
Example 17.1 Milling Hydraulic Systems. The width of a slot cut by a milling machine is important for the proper functioning of a hydraulic system for large tractors. The manufacturer checks the control of the milling process by measuring a sample of five consecutive items during each hour's production. The target width of a slot cut by the milling machine is $\mu = 0.8750$ in. with a target standard deviation of 0.0012 in. What are the centerline and control limits for an s chart? For an $\bar{x}$ chart?

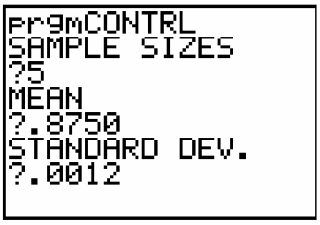
Solution. Bring up the **CONTRL** program and enter either **1** or **2** for the desired variable's control limits. Then enter the sample size of **5** and the target parameters. The centerline and control limits are then displayed.

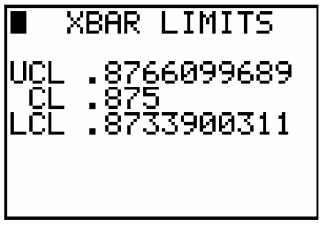












The control limits for the standard deviation are 0 (the standard deviation cannot be negative) and 0.00236. The control limits for the mean are 0.8734 and 0.8766.

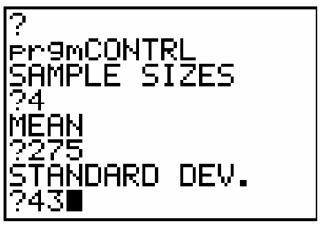
If we have the values of $\bar{x}$ and s from various samples, then we also can use the program to display the control charts. To do so, always enter the values of $\bar{x}$ into list L1 and enter the values of s into list L2. After executing the **CONTRL** program, press ♦ to see the control chart.

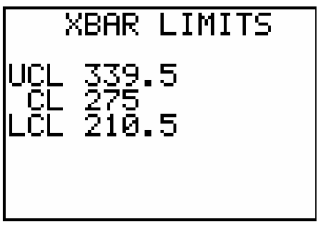
Example 17.2 Computer Monitors. A manufacturer of computer monitors must control the tension on the mesh of fine vertical wires that lies behind the surface of the viewing screen. Too much tension will tear the mesh, and too little will allow wrinkles. The following mesh tension data gives the sample mean and sample deviation from 20 different samples of size 4.

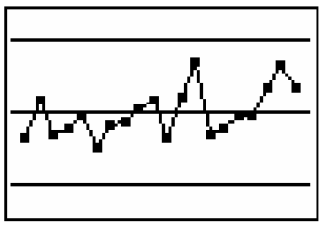
Sample mean	Standard deviation	Sample mean	Standard deviation
253.4	21.8	253.2	16.3
285.4	33.0	287.9	79.7
255.3	45.7	319.5	27.1
260.8	34.4	256.8	21.0
272.7	42.5	261.8	33.0
245.2	42.8	271.5	32.7
265.2	17.0	272.9	25.6
265.6	15.0	297.6	36.5
278.5	44.9	315.7	40.7
285.4	42.5	296.9	38.8

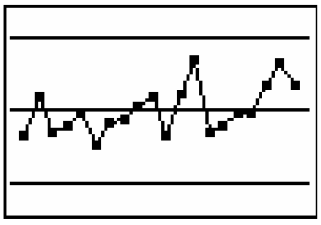
The target mean tension is $\mu = 275$ mV with a target standard deviation of 43 mV. Find the centerline and control limits for $\bar{x}$ and for s. Graph the control charts for each.

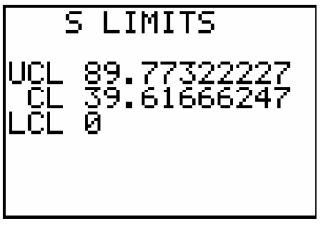
Solution. Enter the sample means into list L1 and the standard deviations into list L2. Then execute the **CONTRL** program for the desired variable to obtain the control limits, and press [GRAPH] to see the control chart. If desired, press [TRACE] and scroll right to see the individual points.

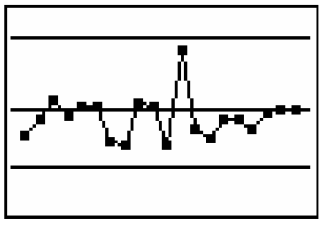












Since both charts are within the control limits, this process is in control; we see only common cause variation around the desired values of both the mean and standard deviation.

17.2 Using Control Charts

We now provide a variation of the **CONTRL** program that will compute the upper and lower control limits and graph the control charts for $\bar{x}$ or s based on past data.

In the program that follows, we note that data sets of *equal* sizes must be entered into lists L1 and L2 in order to obtain the control limits for $\bar{x}$. However, only the values of the sample deviations need to be entered into list L2 in order to compute the control limits for s.

The CONTRL2 Program

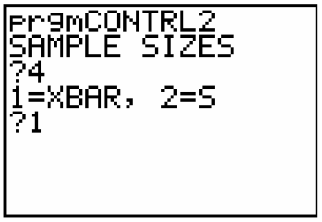
```
Program:CONTRL2
:Disp "SAMPLE SIZES"
:Input N
:If int(N/2)=N/2
:Then
:((N/2-1)!)^2*2^(N-2)/(N-2)!
 *√(2/Π/(N-1))STO▶C
:Else
: √(2Π/(N-1))*(N-1)!/((N-1)/2)!
 /((N-3)/2)!/2^(N-1) STO▶C
:End
:Disp "1=XBAR, 2=S"
:Input W
:If W=1
:Then
:2-Var Stats L1,L2
:x̄STO▶M
:ȳ/CSTO▶S
:"M"STO▶Y1
:"M+3*S/√(N)" STO▶Y2
:"M-3*S/√(N)" STO▶Y3
:seq(I,I,1,dim(L1))STO▶L3
:0STO▶Xmin
:dim(L1)+1STO▶Xmax
:XmaxSTO▶Xscl
:min(min(L1),M-4*S/√(N)) STO▶Ymin
:max(max(L2),M+4*S/√(N)) STO▶Ymax
:Ymax-YminSTO▶Yscl
:Plot1(xyLine,L3,L1,▫)
:ClrHome
:Output(1,4,"XBAR LIMITS")
:Goto 4
:Else
:1-Var Stats L2
: x̄/CSTO▶S
:"C*S"STO▶Y1
:"C*S+3S√(1-C²)"STO▶Y2
:"max(C*S-3S√(1-C²),0)" STO▶Y3
:seq(I,I,1,dim(L2))STO▶L3
:0STO▶Xmin
:dim(L2)+1STO▶Xmax
:XmaxSTO▶Xscl
:min(min(L2),C*S-4S√(1-C²))
STO▶Ymin
:max(max(L2),C*S+4S√(1-C²))
STO▶Ymax
:Ymax-YminSTO▶Yscl
:Plot1(xyLine,L3,L2,▫)
:ClrHome
:Output(1,4,"S LIMITS")
:End
:Lbl 4
:PlotsOff
:PlotsOn 1
:AxesOff
:Output(3,1,"UCL")
:Output(3,5,Y2)
:Output(4,1," CL")
:Output(4,5,Y`)
:Output(5,1,"LCL")
:Output(5,5,Y3)
:Stop
```

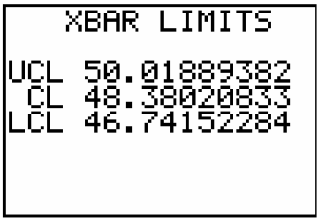
Example 17.3 Measuring Viscosity. The viscosity of a material is its resistance to flow under stress. Viscosity is a critical characteristic of rubber and rubberlike compounds called elastomers, which are used in many consumer products. A specialty chemical company is beginning production of an elastomer that is supposed to have viscosity 45 ± 5 "Mooney Units." The following data gives the mean and standard deviation of elastomer viscosity from samples of size 4 from the first 24 shifts as production begins.

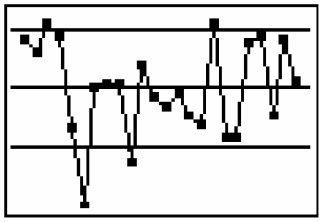
Sample	$\bar{x}$	s	Sample	$\bar{x}$	s
1	49.750	2.684	13	47.875	1.118
2	49.375	0.895	14	48.250	0.895
3	50.250	0.895	15	47.625	0.671
4	49.875	1.118	16	47.375	0.671
5	47.250	0.671	17	50.250	1.566
6	45.000	2.684	18	47.000	0.895
7	48.375	0.671	19	47.000	0.447
8	48.500	0.447	20	49.625	1.118
9	48.500	0.447	21	49.875	0.447
10	46.250	1.566	22	47.625	1.118
11	49.000	0.895	23	49.750	0.671
12	48.125	0.671	24	48.625	0.895

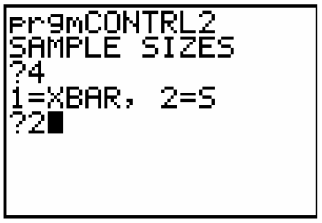
(a) Find the centerline and control limits for $\bar{x}$ and for s based on this past data. Graph the control charts for each.
(b) Remove the two values of s that are out of control and reevaluate the control limits for s based on the remaining data.
(c) Remove the corresponding two values of $\bar{x}$ from the samples that were removed in (b) and reevaluate the control limits for $\bar{x}$ based on the remaining data.

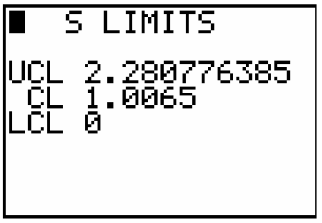
Solution. (a) First, enter the sample means into list L1 and the sample deviations into list L2, then bring up the **CONTRL2** program. Enter 4 for the sample size, then enter **1** when prompted to calculate the control limits for $\bar{x}$ based on this past data. Press GRAPH to see the control chart. Then reexecute the program, but enter **2** when prompted to calculate the control limits for s.

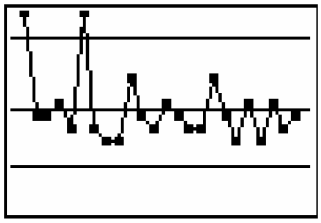






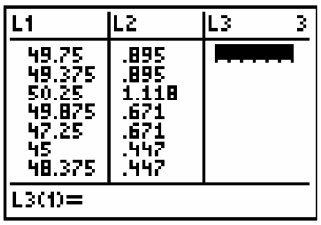


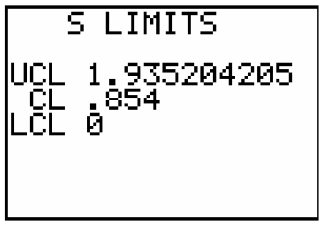


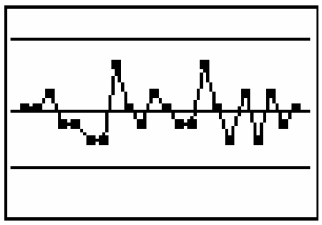


We notice that there were four samples out of control on the $\bar{x}$ chart: the third, sixth, tenth, and seventeenth. There were two samples (the first and sixth) out of control on the *s* chart.

(b) Because the upper control limit for *s* is 2.28, the values of *s* from the first and sixth samples are out of control. We now delete these two values (both 2.684) from list L2 and reexecute the **CONTRL2** program.

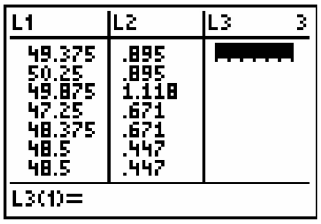


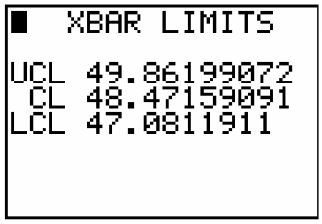


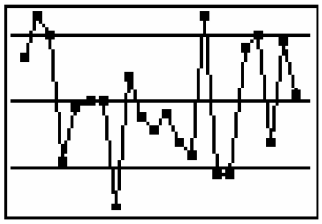


In terms of the standard deviation, this plot is much more desirable.

(c) We now delete the values of $\bar{x}$ from the first and sixth samples from list L1, then reexecute option 1 of the **CONTRL2** program to find the new control limits for $\bar{x}$.







We still see six points outside the control limits, which indicates that this process has not yet reached a controlled state where only common sources of variation are in play.

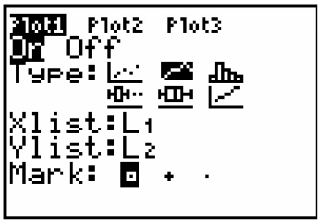
Example 17.4 Joe's Weight. Joe's weight has been stable for several years. An injury keeps Joe away from the gym for several months. The following data give Joe's weight, measured once each week, for the first 16 weeks back at the gym after his injury.

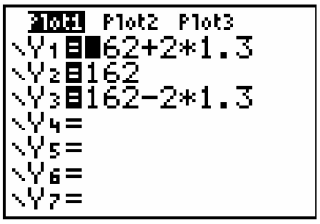
Week	1	2	3	4	5	6	7	8
Weight	168.7	167.6	165.8	167.5	165.3	163.4	163.0	165.5

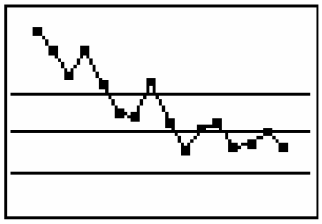
Week	9	10	11	12	13	14	15	16
Weight	162.6	160.8	162.3	162.7	160.9	161.3	162.1	161.0

Joe has a target of μ = 162 pounds for his weight. The short-term variation is estimated to be about σ = 1.3 pounds. Make a control chart for his measurements using control limits $\mu \pm 2\sigma$.

Solution. Simply make a time plot of the measurements together with graphs of the lines $y = \mu$ and $y = \mu \pm 2\sigma$. To do so, enter the weeks into list L1 and the weights into list L2, then adjust the STAT PLOT settings for a scatterplot of L1 versus L2. Next, enter the functions Y1 = 162+2*1.3, Y2 = 162, and Y3 = 162-2*1.3 into the Y= screen, and press ❑✸ to display the graph; adjust the WINDOW settings if needed so that the X range includes all weeks and the Y range includes all weights as well as the upper and lower control limits.







After his injury, Joe's weight was out of control; but it is getting into control.

17.3 Process Capability Indexes

In this section, we provide a program to compute the capability indexes. The program can be used when the parameters are given or when data is given in a list.

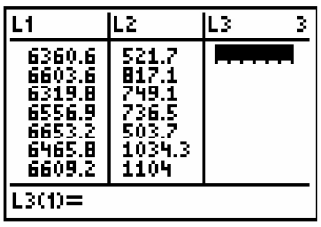
The CAPIND Program

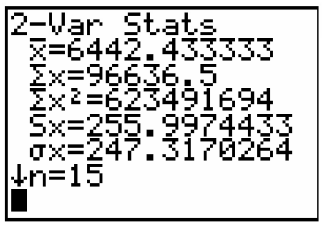
```
PROGRAM:CAPIND
:Disp "LSL"
:Input L
:Disp "USL"
:Input U
:Disp "1=STATS, 2=LIST"
:Input W
:If W=1
:Then
:Disp "MEAN"
:Input M
:Disp "STANDARD DEV."
:Input S
:Goto 1
:Else
:2-Var Stats L1,L2
:x̄ STO▸ M
:ȳ STO▸ S
:End
:Lbl 1
:(U-L)/(6S) STO▸ C
:If M≤L or M≥U
:Then
:0 STO▸ D
:Else
:min(M-L,U-M)/(3S) STO▸ D
:End
:ClrHome
:Output(1,4,"CAP.INDEXES")
:Output(3,2,"Cp")
:Output(3,5,C)
:Output(5,1,"Cpk")
:Output(5,5,D)
:Stop
```

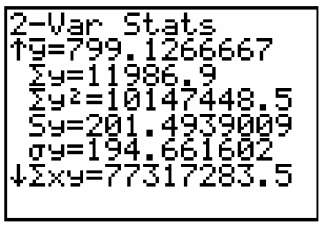
Example 17.5 Hospital Losses. Below are data on a hospital's losses for 120 DRG 209 (major joint replacement) patients collected as 15 monthly samples of eight patients each. The hospital has determined that suitable specification limits for its loss in treating one such patient are LSL = \$4000 and USL = \$8000.
(a) Estimate the percent of losses that meet the specifications.
(b) Estimate C_p and C_{pk}.

Sample	$\bar{x}$	s	Sample	$\bar{x}$	s
1	6360.6	521.7	9	6479.0	704.7
2	6603.6	817.1	10	6175.1	690.5
3	6319.8	749.1	11	6132.4	1128.6
4	6556.9	736.5	12	6237.9	596.6
5	6653.2	503.7	13	6828.0	879.8
6	6465.8	1034.3	14	5925.5	667.8
7	6609.2	1104.0	15	6838.9	819.5
8	6450.6	1033.0			

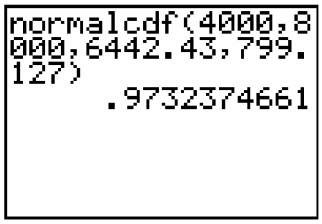
Solution. First, enter the sample means into list L1 and the sample deviations into list L2. Then enter the command 2-Var Stats L1,L2 to compute $\bar{\bar{x}}$ and s. We see that $\bar{\bar{x}}$ = 6442.43 and $\bar{s}$ = 799.127 (from the ȳ output).



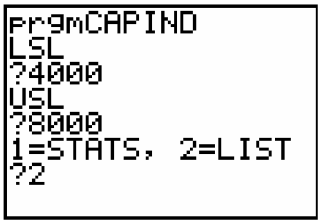


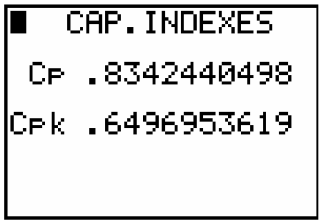


(a) Next, use the normalcdf(command from the DISTR menu to compute $P(4000 \le Z \le 8000)$ for $X \sim N(6442.43, 799.127)$. We find that about 97.32% of losses meet the specifications.



(b) Last, bring up the **CAPIND** program. Enter the LSL of 4000 and the USL of 8000, then enter **2** when prompted for **LIST** to output the capability index approximations.

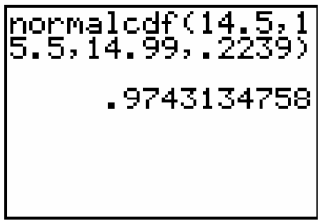




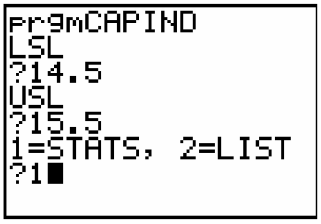
Example 17.6 Measuring Clips. The dimension of the opening of a clip has specifications 15 ± 0.5 millimeters. The production of the clip is monitored by $\bar{x}$ and s charts based on samples of five consecutive clips each hour. The 200 individual measurements from the past week's 40 samples have $\bar{x}$ = 14.99 mm and s = 0.2239 mm.

(a) What percent of clip openings will meet specifications if the process remains in its current state? (b) Estimate the capability index C_{pk}.

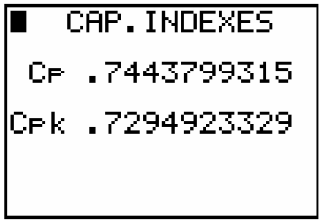
Solution. (a) We must compute $P(15-.5 \le X \le 15+.5)$ for $X \sim N(14.99, .2239)$. Using the `normalcdf(` command, we find that about 97.43% of clip openings will meet specifications if production remains in its current state.



(b) To estimate C_{pk}, we shall use the **CAPIND** program. Bring up the program and enter the LSL of 14.5 and the USL of 15.5, then enter **1** when prompted for **STATS**. Enter the mean and standard deviation to receive the output. We find that C_{pk} = 0.7295.







17.4 Control Charts for Sample Proportions

We conclude this chapter with a program to compute the control limits for sample proportions. The program can be used with summary statistics or when a list of data is given.

If data are in lists, the number of "successes" should be in `L1`, and the number of "trials" should be in `L2`.

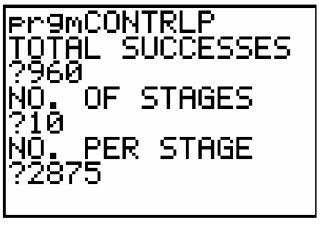
The CONTRLP Program

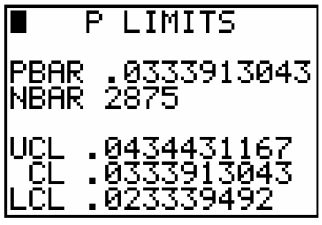
```
PROGRAM:CONTRLP
:Menu("CONTRLP","STATS",1,"LIST",2,"
 QUIT",3)
:Lbl 1
:Disp "TOTAL SUCCESSES"
:Input T
:Disp "NO. OF STAGES"
:Input M
:Disp "NO. PER STAGE"
:Input N
:T/(M*N)→P
:Goto 4
:Lbl 2
:1-Var Stats L2
:x̄→N
:sum(seq(L/(I),I,1,dim(L1)))/sum(
 seq(L2(I),I,1,dim(L2)))→P
:L1/L2→L3
:seq(I,I,1,dim(L1))→L4
:dim(L1)+1→Xmax
:Xmax→Xscl
:min(Y3,min(L3))-.01→Ymin
:max(Y2,max(L3))+.01→Ymax
:1→Yscl
:Plot1(xyLine,L4,L3,▫)
:PlotsOff
:PlotsOn 1
:AxesOff
:Lbl 4
:"P"→Y1
:"min(P+3*√(P(1-P)/N),1)"→Y2
:"max(P-3*√(P(1-P)/N),0)"→Y3
:0→Xmin
:ClrHome
:Output(1,5,"P LIMITS")
:Output(3,1,"PBAR")
:Output(3,6,P)
:Output(4,1,"NBAR")
:Output(4,6,N)
:Output(6,1,"UCL")
:Output(6,5,Y2)
:Output(7,1," CL")
:Output(7,5,Y1)
:Output(8,1,"LCL")
:Output(8,5,Y3)
:Lbl 3
:Stop
```

Example 17.7 Unpaid Invoices. The controller's office of a corporation is concerned that invoices that remain unpaid after 30 days are damaging relations with vendors. To assess the magnitude of the problem, a manager searched payment records for invoices that have arrived in the last 10 months. In that period, an average of 2875 invoices per month have been received, with only 960 remaining unpaid after 30 days. Find $\bar{p}$. Give the centerline and control limits for a p chart.

Solution. Bring up the **CONTRLP** program and press **1** for **STATS**. Enter the values of **960** for total "successes," **10** for the number of "stages," and **2875** for the number per stage. We find that $\bar{p} = 0.0334$ with an LCL of 0.02334 and a UCL of 0.04344.





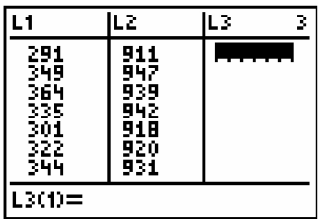


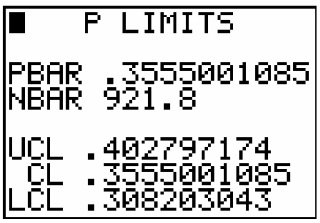
Example 17.8 School Absenteeism. Here are data on the total number of absentees among eighth-graders with three or more unexcused absences at an urban school district. Because the total number of students varies each month, these totals are also given for each month.

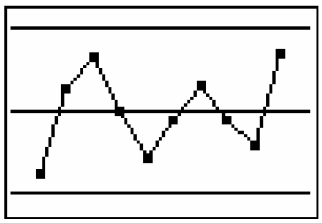
Month	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.
Students	911	947	939	942	918	920	931	925	902	883
Absent	291	349	364	335	301	322	344	324	303	344

(a) Find $\overline{p}$ and $\overline{n}$. (b) Make a p chart using control limits based on $\overline{n}$ students each month.

Solution. (a) First, enter the number of absentees ("successes") for each month into list L1 and the number of students into list L2. Next, execute option 2 of the **CONTRLP** program. We find that $\overline{p} = 0.3555$ and $\overline{n} = 921.8$. (b) After executing the program, the individual monthly proportions are stored in list L3. Press GRAPH to see the p chart that has an LCL of 0.3082 and a UCL of 0.4028.

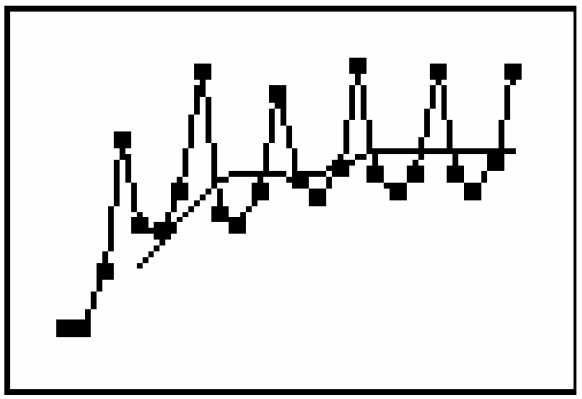






CHAPTER

18



Time Series Forecasting

18.1	Trends and Seasons
18.2	Time Series Models

Introduction

A time series is a sequence of observations on a single variable at equally spaced intervals.

In this chapter, we examine some basic ideas in time series forecasting and modeling. The most basic idea is to use a model to describe a trend (if any), then to add additional terms to describe seasons (or cycles within years). We also examine models in which a new observation is related to a prior ones (autoregressive models) as well as moving average models.

18.1 Trends and Seasons

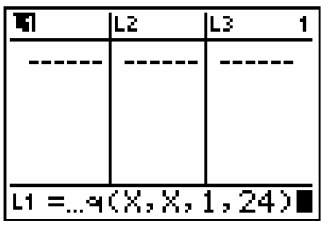
The most basic component of a time series is assessing whether or not there is a trend (systematic rise or fall) and whether there is some aspect that repeats regularly (a cycle, or seasonal component).

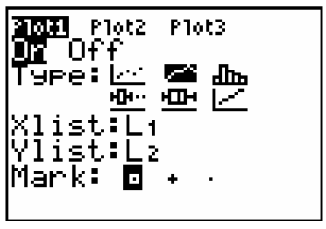
Example 18.1 JCPenney Sales. The table below contains retail sales for JCPenney in millions of dollars beginning with the first quarter of 1996 and ending with the fourth quarter of 2001.

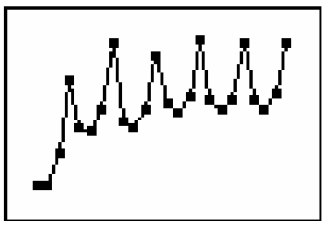
Year-quarter	Sales	Year-quarter	Sales
1996-1st	4452	1999-1st	7339
1996-2nd	4507	1999-2nd	7104
1996-3rd	5537	1999-3rd	7639
1996-4th	8157	1999-4th	9661
1997-1st	6481	2000-1st	7528
1997-2nd	6420	2000-2nd	7207
1997-3rd	7208	2000-3rd	7538
1997-4th	9509	2000-4th	9573
1998-1st	6755	2001-1st	7522
1998-2nd	6483	2001-2nd	7211
1998-3rd	7129	2001-3rd	7729
1998-4th	9072	2001-4th	9542

(a) Make a time plot of the data. Be sure to connect the points in your plot to highlight patterns.
(b) Is there an obvious trend in JCPenney quarterly sales? If so, is the trend positive or negative?
(c) Is there an obvious repeating pattern in this data? If so, clearly describe the repeating pattern.

Solution. (a) First, enter the data into two variables: one (in L1) for the quarter number (from 1 to 24; this is most easily done using the List, Ops, seq(command as shown below) and another for the sales amounts (in L2). Define the plot as a connected scatter plot, then press ZOOM 9 to display it.





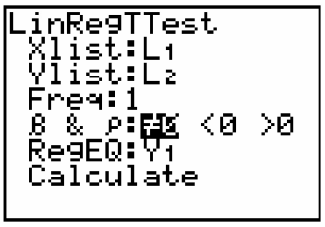


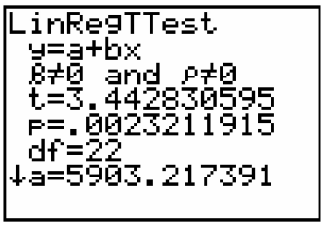
(b) There does appear to be some trend in this plot; however if we eliminate the first couple of years it is not as obvious.

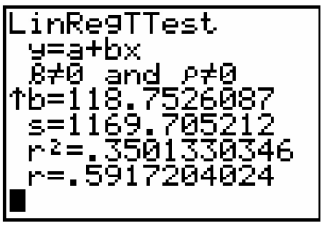
(c) There are definitely repeating patterns in this set of data. The fourth quarter (including Christmas) always has the highest sales; the second quarter always has the lowest sales. Third quarter is generally somewhat higher than the first quarter (back-to-school sales?).

Example 18.2 JCPenney Trends. Find any trend in the JCPenney data using $x = 1$ to correspond with the first quarter of 1996, $x = 2$ to the second quarter of 1996, and so forth. Interpret the slope and intercept in the model.

Solution. Since we already have the data entered, we need to compute the regression line where the dependent variable is Sales (in L2) and the independent variable is the quarter number (in L1). We can use either Stat, Calc, Linreg(a+bx) or Stat, Tests, LinRegTTest to do this.

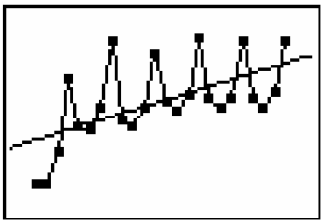






We find the equation *Sales* = 5903.217 + 118.753**Quarter_num*. With a t-statistic of 3.443 and p-value of 0.002, this trend is significantly nonzero. The slope indicates that JCPenney sales increase (on average) 118.753 million dollars per quarter. The intercept says estimated sales for quarter 0 (that would be fourth quarter 1995) are 5903.217 million dollars.

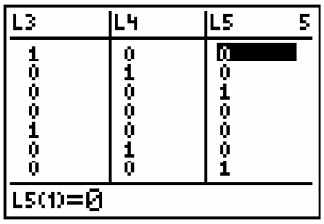
To see the trend line on the time plot, press GRAPH. While significant, the trend does not do a good job of describing the major features of this series; also, most of the trend may be due to the two low values for the first two quarters.



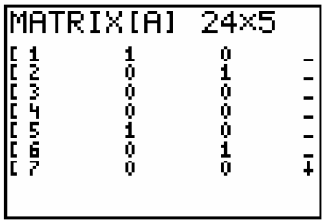
Example 18.3 JCPenney Cycles. Since sales seem to have an annual cycle, add indicator variables for the quarters to the trend-only model fit in Example 18.2. Compare the estimated intercept of this model with the intercept found in Example 18.2. Given the patterns of seasonal variation, which appears to be the better estimate?

Solution. This becomes a multiple linear regression model. From our work in Chapter 11, we know that we need data in a matrix in order to use program **MULREG** to compute the regression. The actual sales data (the dependent variable) must be in the last column of matrix [A]. We will want four predictor variables (the quarter number, and

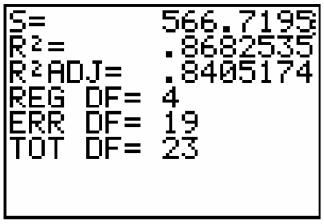
indicator variables for quarters 1, 2, and 3 of each year). If we added a fourth indicator variable we would have linearly dependent columns (because of the intercept) and the model could not be solved. The easiest way to create matrix [A] is to enter the three indicator variable lists, and then use List▸Matrix from the List, Ops menu as shown below.

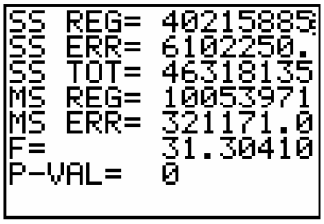


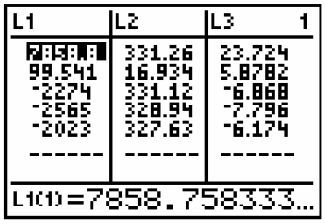




Now, execute program **MULREG** to fit the model (some of the output is omitted).







Our model is now (use the arrow keys to get more decimal places than display on the EDIT screen)

$$Sales = 7858.758 + 99.541 Quarter_num - 2274.21x1 - 2564.585x2 - 2022.792x3.$$

Since the indicator variables only come into play for the selected quarters (as a 1 or 0, which means these values affect the intercept), this really means we have the following set of equations:

$$4thQtrSales = 7858.738 + 99.541 Quarter_num$$
$$1stQtrSales = 5584.448 + 99.541 Quarter_num$$
$$2ndQtrSales = 5294.173 + 99.541 Quarter_num$$
$$3rdQtrSales = 5835.966 + 99.541 Quarter_num$$

Our trend is now that sales increase (on average) 99.541 million dollars per quarter. Since the intercept is still month 0 (fourth quarter 1995), this model forecasts 7858.738 million for that quarter. This is much more reasonable given that fourth quarter sales are always the highest for the year.

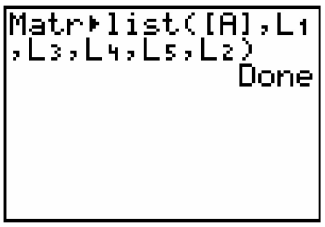
18.2 Time Series Models

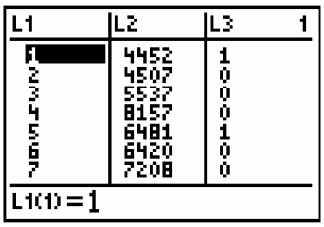
Time series models use past values to predict future values of the series. There are several ways to model these, and we have already looked at regression models. There are several other methods whose study can consume entire statistics courses on their own. We will look at two time series models – the autoregressive model and the moving average model.

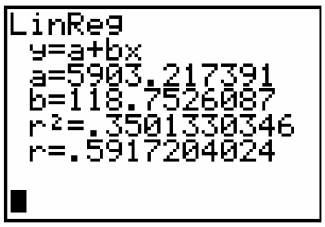
First-order autoregressive models use the equation $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ as a model. This implies that the value one time period prior is the most useful in forecasting the next value. One way to look for an autoregressive relationship is to plot lagged residuals; that is, plot $(\varepsilon_1, \varepsilon_2), (\varepsilon_2, \varepsilon_3), \ldots, (\varepsilon_{n-1}, \varepsilon_n)$.

Example 18.4 JCPenney Trend-Only Residuals. Return to the linear trend only model of Example 18.2 and create a lagged residual plot to examine the data for autocorrelation. Calculate the correlation in the lagged residuals.

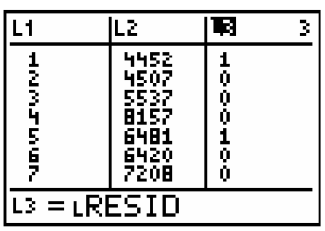
Solution. If you've been working along with this manual, the JCPenney sales data are now in matrix [A] and the original data have been overwritten with the coefficients of the multiple regression model. To reclaim the data (and not have to reenter it), use the command `Matrix▸List` from the `List`, `Ops` menu. In order for us to have the lists we want easily accessible, the actual sales data (in the last column) are specified to end up in `L2`. Recompute the regression; the calculator automatically saves the residuals in a list named `RESID`.

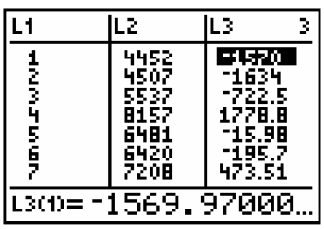


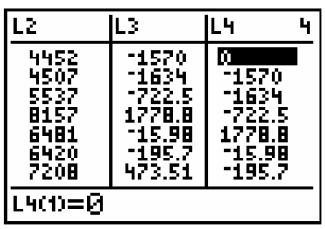




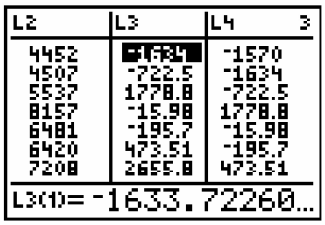
To use this list, move it to `L3`. Highlight the list name, and locate list `RESID` on the `List`, `Names` screen. Press [ENTER] to copy the list into `L3`. To lag the residuals, copy `L3` into `L4` (use `L4=L3` with the `L4` name highlighted on the `EDIT` screen), then [2nd][DEL] to insert a new observation (a 0) as the first entry in `L4`.



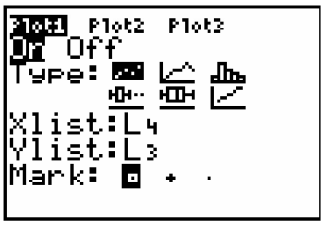


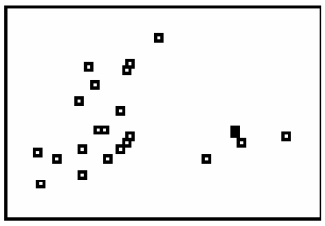


We have one more list manipulation to do before we can create the plot. Adding the new observation to lag the residuals makes the lists unequal in length; also that inserted 0 is not real data. Since calculators must have lists of equal length, delete the first observations in both L3 and L4, and delete the last observation in L4.



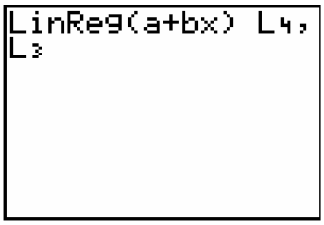
To plot the residuals, define a scatter plot using the lagged residuals (in L4) on the x- axis and the original residuals (in L3) on the y-axis.

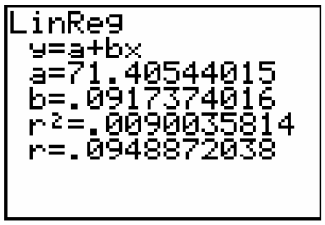




There appear to be two sets of points here: one with a significant amount of correlation in the left part of this plot, and an interesting group with large positive residuals. These are due to fourth quarter sales.

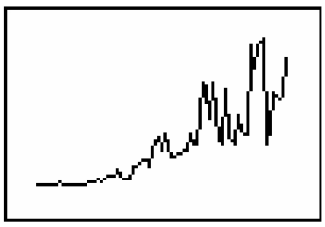
To compute the correlation, simply compute a regression between these two sets of residuals.





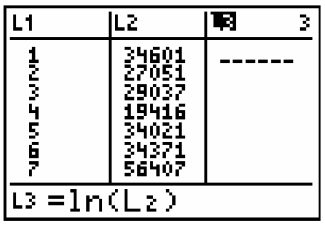
The correlation in these residuals is not significant because the given value of r is 0.095; the linear relationship explains only $r^2 = 0.009$ of the variation in the original residuals. If you use LinRegTTest, you will find the p-value for this relationship is 0.667 which is large. This result is most likely due to the clump of large positive residuals; remember, even a single point can make a correlation seem non-significant.

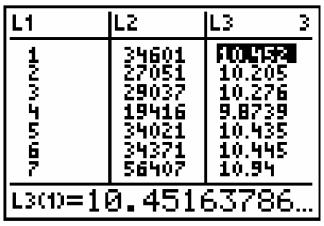
Example 18.5 Autoregressive DVD Sales. The popularity of the DVD format has exploded in the relatively short period of time since its introduction in March 1997. The Consumer Electronics Association tracks monthly sales of DVD players. In the table on the next page are data on DVD sales from April 1997 through June 2002 (63 months). A time plot of the data is shown.

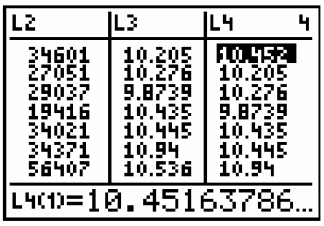


Date	Units sold	Date	Units sold	Date	Units sold
4-97	34601	1-99	125536	10-00	1236658
5-97	27051	2-99	109399	11-00	866507
6-97	29037	3-99	123466	12-00	1303091
7-97	19416	4-99	269107	1-01	572031
8-97	34021	5-99	279756	2-01	555856
9-97	34371	6-99	326668	3-01	1207489
10-97	56407	7-99	325151	4-01	631353
11-97	37657	8-99	260225	5-01	523225
12-97	42575	9-99	501501	6-01	920839
1-98	34027	10-99	603048	7-01	693013
2-98	34236	11-99	449242	8-01	673926
3-98	38336	12-99	646290	9-01	1768821
4-98	42889	1-00	370031	10-01	1516211
5-98	47805	2-00	401035	11-01	1781048
6-98	79044	3-00	412559	12-01	1862772
7-98	84709	4-00	409192	1-02	542698
8-98	81170	5-00	453435	2-02	736118
9-98	113558	6-00	654687	3-02	1162568
10-98	163074	7-00	537453	4-02	1090767
11-98	136908	8-00	557617	5-02	1171984
12-98	233505	9-00	1296280	6-02	1617098

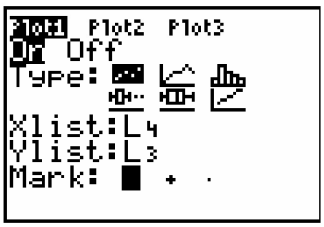
There is very little increase in sales at the beginning; then sales skyrocket and also exhibit considerable variation. As we will see, a scatter plot of lagged ln(DVD sales) versus ln(DVD Sales) is very linear. To first compute ln(DVD sales), with sales in L2, move the cursor to highlight the L3 list name. Enter the command to read L3=ln(L2) as shown below. Press [ENTER] to perform the computation. As we did to lag the residuals in Example 18.4, copy L3 to L4 and insert a 0 at the beginning, then delete the first observation from both L3 and L4, and the last observation from L4. When finished, L3 and L4 should have equal lengths; the first few rows are shown below.

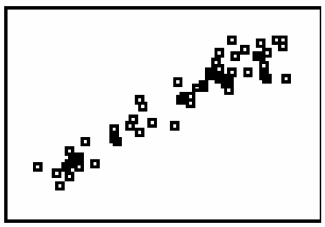




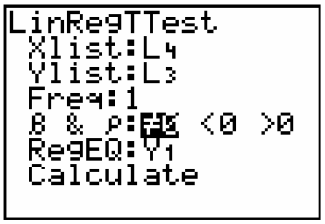


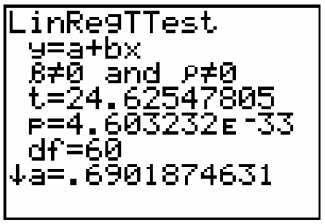
Now, define a scatterplot with L4 on the x axis and L3 on the y axis.

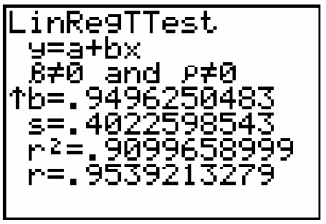




Since this plot is so linear, an autoregressive model appears reasonable. We will fit a linear model to the log sales data using the lagged data as the predictor.







We notice the *t*-statistic for this regression is 24.625 with a *p*-value of essentially 0. This gives an equation of $\ln(DVDsales)_t = .690 + .950\ln(DVDsales)_{t-1}$.

Suppose we want to predict sales for July 2002 based on this model. June 2002 sales were 1617098 units. Putting this into the equation, we have

$$\begin{aligned}\ln(July) &= .690 + .950\ln(1617098)\\ &= .690 + .950 * 14.2961\\ &= 14.271\\ \text{July} &= e^{14.271} = 1577410\end{aligned}$$

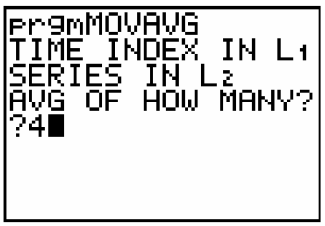
Moving average models forecast y_t as the average of the preceding *k* observations, in other words, $y_t = \frac{y_{t-1} + y_{t-2} + \ldots + y_{t-k}}{k}$. These models smooth out some of the irregular noise in a typical time series. Below, we provide a short program to find the moving averages. The time index must be in L1, and the series in L2. When complete, L3 has the indices for the averaged values and L4 has the averaged values. The program purposefully puts 0s in the first *N* entries in these lists so that one can page through the list and see the moving average estimate alongside the original value; to graph the moving average series along with the original, delete the 0 entries before graphing.

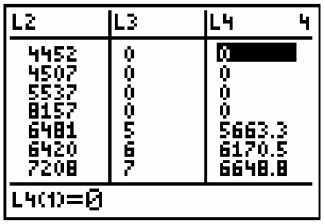
Program MOVAVG

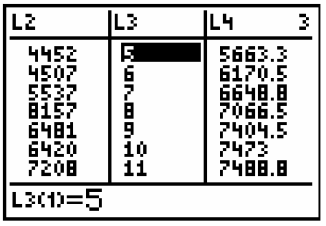
```
Disp "TIME INDEX IN L1"
Disp "SERIES IN L2"
Disp "AVG OF HOW MANY?"
Input N
Clrlist L3, L4
For(I,1,N)
0→L3(I)
0→L4(I)
End
dim(L2)→M
For(K,N+1,M)
K-N→J
sum(seq(L2(L),L,J,J+N-1)/N→L4(K)
L1(K) →L3(K)
End
Stop
```

Example 18.6 JCPenney Moving Averages. JCPenney sales have an annual periodicity. In every year, the fourth quarter sales are highest, and second quarter sales are lowest. Find the moving average predictor using a span of 4. Graph the moving average model on the original time series.

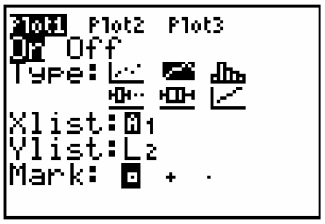
Solution. With the data entered, execute program **MOVAVG**. We want moving averages of length 4. When the program displays the Done message, go to the Statistics Editor. Here, we can see that the first four entries in L3 and L4 are 0s. The next entries are the first two moving averages; 5663.3 = (4453+4507+5537+8157)/4. (More exact values can be found by moving the cursor to the desired entry.) To prepare to graph these series, delete the first four entries in L3 and L4.

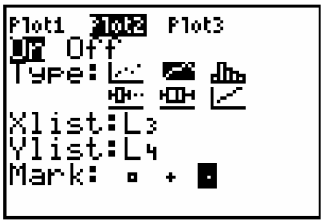


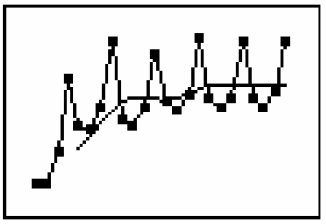




To graph both series at once, define two time plots. Then press ❑✹ to display them. It makes no difference what symbols are plotted for each data point, but we recommend using different ones to help keep the two series differentiated.







Notice that this graph shows the moving average "flattening out" through 2000 and 2001. The country was in a mild recession then, so one would not have expected increasing sales.

Index of Programs

Programs can be downloaded at
http://math.georgiasouthern.edu/~phumphre/TIpgms/ or from your text's portal.

ANOVA1 (page 134) displays the overall sample mean, the pooled sample deviation, the mean square error for groups MSG, the mean square for error MSE, the R^2 coefficient, the F-statistic, and the p-value of the ANOVA test for equality of means when data is entered as summary statistics. Before executing the program, enter the sample sizes into list L1, the sample means into list L2, and the sample deviations into list L3.

ANOVA2W (page 139) displays the p-values for two-way analysis of variance. For one observation per cell, enter the data into matrix [A] before executing the program. For C observations per cell, enter the means into matrix [A] and the standard deviations into matrix [B]. The program also stores the marginal means for the rows and columns into lists L2 and L4. The overall mean is stored as the first value in list L5. The remainder of L5 is the values SSA, SSB, and SSE.

BAYES (page 60) computes the total probability $P(C)$ and conditional probabilities associated with Bayes's rule. Before executing the program, enter values for $P(A_i)$ into list L1 and the conditional probabilities $P(C \mid A_i)$ into list L2. The program displays $P(C)$, stores $P(C \cap A_i)$ in list L3, stores $P(A_i \mid C)$ in list L4, stores $P(A_i \mid C')$ in list L5, and stores $P(C \mid A'_i)$ in list L6.

BOOT (page 145) performs resampling on a random sample in list L1. If a bootstrap confidence interval for the statistic is desired, enter **1** for CONF. INTERVAL? when prompted; otherwise, enter **0**. The program takes resamples from the entered random sample and enters their means into list L2. The mean of all resamples, the bootstrap standard error, and the confidence interval (if specified) are displayed.

BOOTCORR (page 153) performs the bootstrap procedure on the correlation coefficient or the regression slope for paired sample data that has been entered into lists L1 and L2. When prompted, enter **1** if you want to bootstrap the correlation coefficient, or enter **2** if you want to bootstrap the regression slope. The resampled statistics are stored in list L3. The statistic of the original sample data is displayed along with the bootstrap standard error and the confidence interval.

BOOTPAIR (page 150) computes a bootstrap t confidence interval for the difference in means based on random samples that have been entered into lists L1 and L2. The resampled differences in mean are stored in list L3. The difference of the original sample averages is displayed along with the bootstrap standard error and the confidence interval.

BOOTTEST (page 156) performs a permutation test for the difference in means. Before executing, enter data from the first population into list L1 and enter data from the second population into list L2. When prompted, enter **1**, **2**, or **3** to designate the desired alternative. The resampled differences in permuted mean are ordered and then stored in list L3. The program displays the difference between the original sample means and the p-value.

BOOTTRIM (page 149) computes a bootstrap t confidence interval for a trimmed mean on a random sample that has been entered into list L1. When prompted, enter the desired number of

resamples, the decimal amount to be trimmed at each end, and the desired confidence level. The program takes resamples from the entered random sample and enters their trimmed means into list L2. The trimmed mean of the original sample, the trimmed bootstrap standard error, and the confidence interval are displayed.

BTPRTEST (page 158) performs a permutation test for either the difference in paired means or for the correlation. Before executing, enter the data set into lists L1 and L2. When prompted, enter **1** or **2** to designate the desired test, then enter **1**, **2**, or **3** to designate the desired alternative. The resampled permuted pair differences in mean (or correlations) are ordered and stored in list L3. The statistic from the original paired sample is displayed along with the *p*-value.

CAPIND (page 186) computes capability indexes based on given parameters or on data that has been entered into lists L1 and L2.

CONTRAST (page 136) computes the *p*-value for a significance test and a confidence interval for mean population contrasts. Before executing the program, enter the sample sizes into list L1, the sample means into list L2, the sample standard deviations into list L3, and the contrast equation coefficients into list L4. When prompted during the program, enter either **1** or **2** for a one-sided or two-sided alternative.

CONTRL (page 180) computes the upper and lower control limits and graphs the control charts for $\overline{x}$ and s.

CONTRL2 (page 183) computes the upper and lower control limits and graphs the control charts for $\overline{x}$ or s based on past data. Data sets of equal sizes must be entered into lists L1 and L2 in order to obtain the control limits for $\overline{x}$. Only the values of the sample deviations need to be entered into list L2 in order to compute the control limits for s.

CONTRLP (page 189) computes the control limits for sample proportions given either summary statistics or data entered into lists L1 and L2.

FITTEST (page 108) performs a goodness of fit test for a specified discrete distribution. Before executing, enter the specified proportions into list L1 and enter the observed cell counts into list L2. The expected cell counts are computed and stored in list L3; the contributions to the chi-square test statistic are stored in list L4. The program displays the test statistic and the *p*-value.

KRUSKAL (page 169) performs the Kruskal-Wallace test. Before executing, enter the data into the columns of matrix [A] and the sample sizes into a row matrix [B]. The program displays the test statistic and *p*-value. List L3 will contain the merged, sorted measurements, L4 will contain their (averaged) ranks, and L5 will contain the sum of ranks from each population.

LOG1 (page 175) computes and displays the coefficients of the linear regression model for the log of odds ratio $\ln(p/(1-p)) = \beta_0 + \beta_1 x$. Also displays the odds ratio.

MOVAVG (page 198) calculates the moving average of N values for a time series in L2 with time indices in L1. The time indices corresponding to the moving averages are stored in L3 and the averages in L4.

MULREG (page 123) computes the regression coefficients and an ANOVA table for the multiple linear regression model $\mu_Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. The squared correlation

coefficient, F-statistic, p-value, and the standard deviation are also displayed. Before executing the program, enter sample data as columns into matrix [A] with the last column used for the dependent variable.

ODDS (page 173) computes the appropriate mathematical odds for a given probability p of an event A. If $p \le 0.50$, then the odds against A are given as the ratio $(1-p):p$. If $p > 0.50$, then the odds in favor of A are given as the ratio $p:(1-p)$.

ODDS2 (page 174) computes the odds-in-favor ratio between two proportions.

ODDSINT (page 176) computes a confidence interval for the slope β_1 of the logistic regression model and the odds ratio.

ODDSTEST (page 176) computes the test statistic and p-value for the hypothesis test that an odds ratio equals 1.

PSAMPSZE (page 95) computes the required sample size that would give a maximum desired margin of error m for a confidence interval about a proportion.

RANDOM (page 46) randomly chooses a subset of specified size from the set {1, 2, . . ., n} and stores the values in list L1.

RANKSUM (page 163) performs the Wilcoxon rank sum test on data from two populations entered into lists L1 and L2. The program displays the expected sum of ranks from list L1, the sums of the ranks from each list, and the smallest tail value created by the test statistic which is the sum of the ranks from L1. List L3 then contains the merged, sorted measurements, and L4 contains their (averaged) ranks.

REGANV (page 118) computes the ANOVA table for linear regression and displays the associated F-statistic and p-value. Before executing the program, data must be entered into lists L1 and L2. The ANOVA table is stored into lists L4, L5, and L6.

REGINF (page 121) computes confidence intervals for the slope and for the intercept of the linear regression model $y = \beta_0 + \beta_1 x$, as well as confidence intervals for the mean value of y at a given x and prediction intervals y given a new value of x.

SIGNRANK (page 166) performs the Wilcoxon signed rank test on data sets of size n from two populations. Before executing, enter the data into lists L1 and L2. The program sorts the absolute value of the differences L2 – L1 into list L3, but disregards any zero differences. The (averaged) rank of each nonzero difference is stored in list L4. The sums of the ranks of the positive and negative differences are displayed. The program also displays the smallest tail value created by the test statistic, which is the sum of the ranks of the positive differences.

TSCORE (page 82) finds the critical value t^* of a t distribution upon specifying the degrees of freedom and confidence level.

ZSAMPSZE (page 75) computes the sample size needed to obtain a desired maximum margin of error with a specified level of confidence when finding a confidence interval for the mean using a known standard deviation and normal distribution z-scores.

Chapter 1 Exercises

1.7 Refer to the first exam scores from Exercise 1.5 (reproduced below) and this histogram you produced in Exercise 1.6. Now make a histogram for these data using classes 40 – 59, 60 – 79, and 80 – 100. Compare this histogram with the one that you produced in Exercise 1.6.

80	73	92	85	75	98	93	55	80	90	92	80	87	90	72
65	70	85	83	60	70	90	75	75	58	68	85	78	80	93

1.19 Email spam is the curse of the Internet. Here is a compilation of the most common types of spam:

Type of spam	Percent
Adult	14.5
Financial	16.2
Health	7.3
Leisure	7.8
Products	21.0
Scams	14.2

Make two bar graphs of these percents, one with bars ordered as in the table (alphabetical, and the other with bars in order from tallest to shortest. Comparisons are easier if you order the bars by height. A bar graph ordered from tallest to shortest is sometimes called a **Pareto chart**, after the Italian economist who recommended this procedure.

1.31 Table 1.7 (reproduced below) contains data on the mean annual temperatures (degrees Fahrenheit) for the years 1941 to 2000 at two locations in California: Pasadena and Redding. Make time plots of both time series and compare their main features. You can see why discussions of climate change often bring disagreement.

Year	Pasadena	Redding	Year	Pasadena	Redding
1951	62.27	62.02	1976	64.23	63.51
1952	61.59	62.27	1977	64.47	63.89
1953	62.64	62.06	1978	64.21	64.05
1954	62.88	61.65	1979	63.76	60.38
1955	61.75	62.48	1980	65.02	60.04
1956	62.93	63.17	1981	65.80	61.95
1957	63.72	62.42	1982	63.50	59.14
1958	65.02	64.42	1983	64.19	60.66
1959	65.69	65.04	1984	66.06	61.72

1960	64.48	63.07	1985	64.44	60.51
1961	64.12	63.50	1986	65.31	61.76
1962	62.82	63.97	1987	64.58	62.94
1963	63.71	62.42	1988	65.22	63.70
1964	62.76	63.29	1989	64.53	61.50
1965	63.03	63.32	1990	64.96	62.22
1966	64.25	64.51	1991	65.60	62.73
1967	64.36	64.21	1992	66.07	63.59
1968	64.15	63.40	1993	65.16	61.55
1969	63.51	63.77	1994	64.63	61.63
1970	64.08	64.30	1995	65.43	62.62
1971	63.59	62.23	1996	65.76	62.93
1972	64.53	63.06	1997	66.72	62.48
1973	63.46	63.75	1998	64.12	60.23
1974	63.93	63.80	1999	64.85	61.88
1975	62.36	62.66	2000	66.25	61.58

1.47 Here are the scores on the first exam in an introductory statistics course for 10 students.

80	73	92	85	75	98	93	55	80	90

Find the mean first exam score for these students.

1.49 Here are the scores on the first exam in an introductory statistics course for 10 students.

80	73	92	85	75	98	93	55	80	90

Find the quartiles for these first-exam scores.

1.51 Here are the scores on the first exam in an introductory statistics course for 10 students.

80	73	92	85	75	98	93	55	80	90

Make a boxplot for these first-exam scores.

1.57 C-reactive protein (CRP) is a substance that can be measured in the blood. Values increase substantially within 6 hours of an infection and reach a peak within 24 to 48 hours after. In adults, chronically high values have been linked to an increased risk of

cardiovascular disease. In a study of apparently healthy children aged 6 to 60 months in Papua, New Guinea, CRP was measured in 90 children. The units are milligrams per liter (mg/l). Here are the data from a random sample of 40 of these children:

0.00	0.00	30.61	46.70	22.82	0.00	5.36	59.76	0.00	20.78
3.90	5.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.10
5.64	3.92	73.20	0.00	0.00	4.81	5.66	15.74	0.00	7.89
8.22	6.81	0.00	26.41	3.49	9.57	0.00	0.00	9.37	5.53

(a) Find the five-number summary for these data.
(b) Make a boxplot.
(c) Make a histogram.
(d) Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

1.103 Consider the ISTEP scores, which are approximately Normal, $N(572, 51)$. Find the proportion of students who have scores less than 600. Find the proportion of students who have scores greater than or equal to 600. Sketch the relationship between these two calculations using pictures of Normal curves similar to the ones given in Example 1.27.

1.123. The variable Z has a standard Normal distribution.
(a) Find the number z that has cumulative proportion 0.85.
(b) Find the number z such that the event $Z > z$ has proportion 0.40.

1.131 Reports on a student's ACT or SAT usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent; the percent of all scores that were lower than this one. Jacob scores 16 on the ACT. What is his percentile?

1.139 The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal with mean 266 days and standard deviation 16 days.
(a) What percent of pregnancies last less than 240 days (that's about 8 months)?
(b) What percent of pregnancies last between 240 and 270 days (roughly between 8 months and 9 months)?
(c) How long do the longest 20% of pregnancies last?

1.147 We expect repeated careful measurements of the same quantity be be approximately Normal. Make a Normal quantile plot for Cavendish's measurements in Exercise 1.40 (data reproduced below). Are the data approximately Normal? If not, describe any clear deviations from Normality.

5.50	5.55	5.57	5.34	5.42	5.30
5.61	5.36	5.53	5.79	5.47	5.75
4.88	5.29	5.62	5.10	5.63	5.68
5.07	5.58	5.29	5.27	5.34	5.85
5.26	5.65	5.44	5.39	5.46	

Chapter 2 Exercises

2.7 Here are the data for the second test and the final exam for the same students as in Exercise 2.6:

Second-test score	158	163	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

(a) Explain why you should use the second-test score as the explanatory variable.
(b) Make a scatterplot and describe the relationship.
(c) Why do you think the relationship between the second-test score and the final-exam score is stronger than the relationship between the first-test score and the final-exam score?

2.21 Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The table below gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.

(a) Make a scatterplot of the data, using different symbols or colors for men and women.
(b) Is the association between these variables positive or negative? How strong is the relationship? Does the pattern of the relationship differ for women and men? How do the male subjects as a group differ from the female subjects as a group?

Sex	Mass	Rate	Sex	Mass	Rate
M	62.0	1792	F	40.3	1189
M	62.9	1666	F	33.1	913
F	36.1	995	M	51.9	1460
F	54.6	1425	F	42.4	1124
F	48.5	1396	F	34.5	1052
F	42.0	1418	F	51.1	1347
M	47.4	1362	F	41.2	1204
F	50.6	1502	M	51.9	1867
F	42.0	1256	M	46.9	1439
M	48.7	1614			

2.23 Table 2.3 (reproduced below) shows the progress of world record times (in seconds) for the 10,000 meter run up to mid-2004. Concentrate on the women's world record times. Make a scatterplot with year as the explanatory variable. Describe the pattern of improvement over time that your plot displays.

Women's Record Times			
1967	2286.4	1982	1895.3
1970	2130.5	1983	1895.0
1975	2100.4	1983	1887.6
1975	2041.4	1984	1873.8
1977	1995.1	1985	1859.4
1979	1972.5	1986	1813.7
1981	1950.8	1993	1771.8
1981	1937.2		

2.31 Here are the data for the second test and the final exam for the same students as in Exercise 2.6 (and 2.30):

Second-test score	158	163	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

Find the correlation between these two variables.

2.45 Table 1.10 (reproduced below) gives the city and highway gas mileage for 21 two-seater cars, including the Honda Insight gas-electric hybrid car.

(a) Make a scatterplot of highway mileage y against city mileage x for all 21 cars. There is a strong positive linear association. The Insight lies far from the other points. Does the Insight extend the linear pattern of the other card, or is it far from the line they form?

(b) Find the correlation between city and highway mileages both without and with the Insight. Based on your answer to (a), explain why r changes in this direction when you add the Insight.

City	Hwy	City	Hwy
17	24	9	13
20	28	15	22
20	28	12	17
17	25	22	28
18	25	16	23
12	20	13	19

11	16	20	26
10	16	20	29
17	23	15	23
60	66	26	32
9	15		

2.59 Here are the data for the second test and the final-exam scores (again).

Second-test score	158	163	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

(a) Plot the data with the second-test scores on the x axis and the final-exam scores on the y axis.
(b) Find the least-squares regression line for predicting the final-exam score using the second-test score.
(c) Graph the least-squares regression line on your plot.

2.69 Table 2.4 (reproduced below) gives data on the growth of icicles at two rates of water flow. You examined these data in Exercise 2.24. Use least-squares regression to estimate the rate (centimeters per minute) at which icicles grow at these two flow rates. How does flow rate affect growth?

Run 8903				**Run 8905**			
Time (min)	**Length (cm)**	**Time(min)**	**Length(cm)**	**Time(min)**	**Length(cm)**	**Time (min)**	**Length (cm)**
10	0.6	130	18.1	10	0.3	130	10.4
20	1.8	140	19.9	20	0.6	140	11.0
30	2.9	150	21.0	30	1.0	150	11.9
40	4.0	160	23.4	40	1.3	160	12.7
50	5.0	170	24.7	50	3.2	170	13.9
60	6.1	180	27.8	60	4.0	180	14.6
70	7.9			70	5.3	190	15.8
80	10.1			80	6.0	200	16.2
90	10.9			90	6.9	210	17.9
100	12.7			100	7.8	220	18.8
110	14.4			110	8.3	230	19.9
120	16.6			120	9.6	240	21.1

2.87 A study of nutrition in developing countries collected data from the Egyptian village of Nahya. Here are the mean weights (in kilograms) for 170 infants in Nahya who were weighed each month during their first year of life:

Age (months)	1	2	3	4	5	6	7	8	9	10	11	12
Weight (kg)	4.3	5.1	5.7	6.3	6.8	7.1	7.2	7.2	7.2	7.2	7.5	7.8

(a) Plot weight against time.

(b) A hasty user of statistics enters the data into software and computes the least-squares line without plotting the data. The result is

```
The regression equation is
Weight = 4.88 + 0.267 age
```

Plot this line on your graph. Is it an acceptable summary of the overall pattern of growth? Remember that you can calculate the least-squares line for *any* set of two-variable data. It's up to you to decide if it makes sense to fit a line.

(c) Fortunately, the software also prints out the residuals from the least-squares line. In order of age along the rows, they are

–0.85	–0.31	0.02	0.35	0.58	0.62
0.45	0.18	–0.08	–0.35	–0.32	–0.28

Verify that the residuals have sum zero (except for roundoff error). Plot the residuals against age and add a horizontal line at zero. Describe carefully the pattern that you see.

2.93 Careful statistical studies often include examination of potential lurking variables. This was true of the study of the effect of nonexercise activity (NEA) on fat gain (Example 2.12, page 109), our lead example in Section 2.3. Overeating may lead our bodies to spontaneously increase NEA (fidgeting and the like). Our bodies might also spontaneously increase their basal metabolic rate (BMR), which measures energy use while resting. If both energy uses increased, regressing fat gain on NEA alone would be misleading. Here are data on BMR and fat gain for the same 16 subjects whose NEA we examined earlier:

BMR increase (cal)	117	352	244	–42	–3	134	136	–32
Fat gain (kg)	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
BMR increase (cal)	–99	9	–15	–70	165	172	100	35
Fat gain (kg)	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

The correlation between NEA and fat gain is $r = -0.7786$. The slope of the regression line for predicting fat gain from NEA is $b_1 = -0.00344$ kilogram per calorie. What are the correlation and slope for BMR and fat gain? Explain why these values show that BMR has much less effect on fat gain than does NEA.

2.119 A market research firm conducted a survey of companies in its state. They mailed a questionnaire to 300 small companies, 300 medium-sized companies, and 300 large companies. The rate of nonresponse is important in deciding how reliable survey results are. Here are the data on response to this survey.

Size of company	Response	No response	Total
Small	175	125	300
Medium	145	155	300
Large	120	180	300

(a) What is the overall percent of nonresponse?
(b) Describe how nonresponse is related to the size of business. (Use percents to make your statements precise.)
(c) Draw a bar graph to compare the nonresponse percents for the three size categories.
(d) Using the total number of responses as a base, compute the percent of responses that come from each of small, medium, and large businesses.
(e) The sampling plan was designed to obtain equal numbers of responses from small, medium, and large companies. In preparing an analysis of the survey results, do you think it would be reasonable to proceed as if the responses represented companies of each size equally?

Chapter 3 Exercises

3.27 Doctors identify "chronic tension-type headaches" as headaches that occur almost daily for at least six months. Can antidepressant medications or stress management training reduce the number and severity of these headaches? Are both together more effective than either alone? Investigators compared four treatments: antidepressant alone, placebo alone, antidepressant plus stress management, and placebo plus stress management. Outline the design of the experiment. The headache sufferers named below have agreed to participate in the study. Use software or Table B at line 151 to randomly assign the subjects to the treatments.

Anderson	Archberger	Bezawada	Cetin	Cheng
Chronopoulou	Codrington	Daggy	Daye	Engelbrecht
Guha	Hatfield	Hua	Kim	Kumar
Leaf	Li	Lipka	Lu	Martin
Mehta	Mi	Nolan	Olbricht	Park
Paul	Rau	Saygin	Shu	Tang
Towers	Tyner	Vassilev	Wang	Watkins
Xu				

3.43 We often see players on the sidelines of a football game inhaling oxygen. Their coaches think this will speed their recovery. We might measure recovery from intense exercise as follows: Have a football player run 100 yards three times in quick succession. Then allow three minutes to rest before running 100 yards again. Time the final run. Because players vary greatly in speed, you plan a matched pairs experiment using 20 football players as subjects. Describe the design of such an experiment to investigate the effect of inhaling oxygen during the rest period. Why should each player's two trials be on different days? Use Table B at line 140 to decide which players will get oxygen on their first trial.

3.51 The walk to your statistics class takes about 10 minutes, about the amount of time needed t listen to three songs on your iPod. You decide to take a simple random sample of songs from a Billboard list of Rock Songs. Here is the list:

1	Miss Murder	2	Animal I Have Become	3	Steady As She Goes	4	Dani California
5	The Kill (Bury Me)	6	Original Fire	7	When You Were Young	8	MakeD – Sure
9	Vicarious	10	The Diary of Jane				

Select the three songs for your iPod using a simple random sample.

3.57 You are planning a report on apartment living in a college town. You decide to select 5 apartment complexes at random for in-depth interviews with residents. Select a simple random sample of 5 of the following apartment complexes. If you use Table B, start at line 137.

1	Ashley Oaks	2	Country View	3	Mayfair Village
4	Bay Pointe	5	Country Villa	6	Nobb Hill
7	Beau Jardin	8	Crestview	9	Pemberly Courts
10	Bluffs	11	Del-Lynn	12	Peppermill
13	Brandon Place	14	Fairington	15	Pheasant Run
16	Briarwood	17	Fairway Knolls	18	Richfield
19	Brownstone	20	Fowler	21	Sagamore Ridge
22	Burberry	23	Franklin Park	24	Salem Courthouse
25	Cambridge	26	Georgetown	27	Village Manor
28	Chauncey Village	29	Greenacres	30	Waterford Court
31	Country Squire	32	Lahr House	33	Williamsburg

3.67 Stratified samples are widely used to study large areas of forest. Based on satellite images, a forest area in the Amazon basin is divided into 14 types. Foresters studied the four most commercially valuable types: alluvial climax forests of quality levels 1, 2, and 3, and mature secondary forest. They divided the area of each type into large parcels, chose parcels of each type at random, and counted tree species in a 20-by-25 meter rectangle randomly placed within each parcel selected. Here is some detail:

Forest type	**Total parcels**	**Sample size**
Climax 1	36	4
Climax 2	72	7
Climax 3	31	3
Secondary	42	4

Choose the stratified sample of 18 parcels. Be sure to explain how you assigned labels to parcels. If you use Table B, start at line 140.

3.91 We can construct a sampling distribution by hand in the case of a very small population. The population contains 10 students. Here are their scores on an exam:

Student	0	1	2	3	4	5	6	7	8	9
Score	82	62	80	58	72	73	65	66	74	62

The parameter of interest is the mean score, which is 69.4. The sample is an SRS of $n = 4$ students drawn from this population. The students are labeled 0 to 9 so that a simple random digit from table B chooses one student for the sample.

(a) Use table B to draw an SRS of size 4 from this population. Write the four scores in your sample and calculate the mean $\bar{x}$ of the sample scores. This statistic is an estimate of the population parameter.

(b) Repeat this process 9 more times. Make a histogram of the 10 values of $\bar{x}$. Is the center of your histogram close to 69.4? (Ten repetitions give only a crude approximation to the sampling distribution. If possible, pool your work with that of other students – using different parts of Table B – to obtain several hundred repetitions and make a histogram of the values of $\bar{x}$. This histogram is a better approximation to the sampling distribution.)

Chapter 4 Exercises

4.7 The basketball player Shaquille O'Neal makes about half of his free throws over an entire season. Use Table B or the *Probability* applet to simulate 100 free throws shot independently by a player who has probability 0.5 of making each shot.

(a) What percent of the 100 shots did he hit?

(b) Examine the sequence of hits and misses. How long was the longest run of shots made? Of shots missed? (Sequences of random outcomes often show runs longer than our intuition thinks likely.)

4.51 Spell-checking software catches "nonword errors," which result in a string of letters that is not a word, as when "the" is typed as "teh." When undergraduates are asked to write a 250 word essay (without spell checking), the number X of nonword errors has the following distribution:

Value of X	0	1	2	3	4
Probability	0.1	0.3	0.3	0.2	0.1

Sketch the probability distribution for this random variable.

4.65 How many close friends do you have? Suppose that the number of close friends adults claim to have varies from person to person with mean $\mu = 9$ and standard deviation $\sigma = 2.5$. An opinion poll asks this question of an SRS of 1100 adults. We will see in the next chapter that in this situation the sample mean response $\bar{x}$ has approximately the Normal distribution with mean 9 and standard deviation 0.075. What is $P(8 \le \bar{x} \le 10)$, the probability that the statistic $\bar{x}$ estimates the parameter μ to within ±1?

4.73 Example 4.22 gives the distribution of grades (A = 4, B = 3, and so on)in English 210 at North Carolina State University as

Value of X	0	1	2	3	4
Probability	0.05	0.04	0.20	0.40	0.31

Find the average (that is, the mean) grade in this course.

4.89 According to the current Commissioners' Standard Ordinary mortality table, adopted by state insurance regulators in December 2002, a 25-year-old man has these probabilities of dying during the next five years:

Age at death	25	26	27	28	29
Probability	0.00039	0.00044	0.00051	0.00057	0.00060

(a) What is the probability that the man does not die in the next five years?
(b) An online insurance site offers a term insurance policy that will pay $100,000 if a 25-year-old man dies within the next 5 years. The cost is $175 per year. So the insurance company will take in $875 from this policy if the man does not die within five years. If he does die, the company must pay $100,000. Its loss depends on how many premiums were paid, as follows:

Age at death	25	26	27	28	29
Loss	$99,825	$99,650	$99,475	$99,300	$99,125

What is the insurance company's mean cash intake from such policies?

4.137 A grocery store gives its customers cards that may win them a prize when matched with other cards. The back of the card announces the following probabilities of winning various amounts if a customer visits the store 10 times:

Amount	$1000	$250	$100	$10
Probability	1/10,000	1/1000	1/100	1/20

(a) What is the probability of winning nothing?
(b) What is the mean amount won?
(c) What is the standard deviation of the amount won?

Chapter 5 Exercises

5.5 (a) Suppose X has the $B(4, 0.3)$ distribution. Use software or Table C to find $P(X = 0)$ and $P(X \geq 3)$.

(b) Suppose X has the $B(4, 0.7)$ distribution. Use software or Table C to find $P(X = 4)$ and $P(X \leq 1)$.

5.7 Suppose we toss a fair coin 100 times. Use the Normal approximation to find the probability that the sample proportion is

(a) between 0.4 and 0.6.

(b) between 0.45 and 0.55.

5.13 Typographic errors in a text are either nonword errors (as when "the" is typed as "teh") or word errors that result in a real but incorrect word. Spell-checking software will catch nonword errors but not word errors. Human proofreaders catch 70% of word errors. You ask a fellow student to proofread an essay in which you have deliberately made 10 word errors.

(a) If the student matches the usual 70% rate, what is the distribution of the number of errors caught? What is the distribution of the number of errors missed?

(b) Missing 4 or more out of 10 errors seems a poor performance. What is the probability that a proofreader who catches 70% of word errors misses 4 or more out of 10?

5.17 In the proofreading setting of Exercise 5.13, what is the smallest number of misses m with $P(X \geq m)$ no larger than 0.05? You might consider m or more misses as evidence that a proofreader actually catches fewer than 70% of word errors.

5.21 Children inherit their blood type from their parents, with probabilities that reflect the parents' genetic makeup. Children of Juan and Maria each have probability 1/4 of having blood type A and inherit independently of each other. Juan and Maria plan to have 4 children; let X be the number who have blood type A.

(a) What are n and p in the binomial distribution of X?

(b) Find the probability of each possible value of X, and draw a probability histogram for this distribution.

(c) Find the mean number of children with type A blood, and mark the location of the mean on your probability histogram.

5.25 The Harvard College Alcohol Study finds that 67% of college students support efforts to "crack down on underage drinking." The study took a sample of almost 15,000 students, so the population proportion who support a crackdown is very close to

p = 0.67. The administration of your college surveys an SRS of 200 students and finds that 140 support a crackdown on underage drinking.

(a) What is the sample proportion who support a crackdown on underage drinking?

(b) If in fact the proportion of all students on your campus who support a crackdown is the same as the national 67%, what is the probability that the proportion in an SRS of 200 students is as large or larger than the result of the administration's sample?

5.31 One way of checking the effect of undercoverage, nonresponse, and other sources of error in a sample survey is to compare the sample with known demographic facts about the population. The 2000 census found that 23,772,494 of the 209,128,094 adults (aged 18 and over) in the United States called themselves "Black or African American."

(a) What is the population proportion p of blacks among American adults?

(b) An opinion poll chooses 1200 adults at random. What is the mean number of blacks in such samples? (Explain the reasoning behind your calculation.)

(c) Use a Normal approximation to find the probability that such a sample will contain 100 or fewer blacks. Be sure to check that you can safely use the approximation.

5.49 The gypsy moth is a serious threat to oak and aspen trees. A state agriculture department places traps throughout the state to detect the moths. When traps are checked periodically, the mean number of moths trapped is only 0.5, but some traps have several moths. The distribution of moth counts is discrete and strongly skewed, with standard deviation 0.7.

(a) What are the mean and standard deviation of the average number of moths $\bar{x}$ in 50 traps?

(b) Use the central limit theorem to find the probability that the average number of moths in 50 traps is greater than 0.6.

5.53 Sheila's measured glucose level one hour after ingesting a sugary drink varies according to the Normal distribution with μ =125 mg/dl and σ = 10 mg/dl. What is the level L such that there is probability only 0.05 that the mean glucose level of 3 test results falls above L for Sheila's glucose level distribution?

5.57 The distribution of annual returns on common stocks is roughly symmetric, but extreme observations are more frequent than in a Normal distribution. Because the distribution is not strongly non-Normal, the mean return over even a moderate number of years is close to Normal. Annual real returns on the Standard & Poor's 500 stock index over the period 1871 to 2004 have varied with mean 9.2% and standard deviation 20.6%. Andrew plans to retire in 45 years and is considering investing in stocks. What is the probability (assuming that the past pattern of variation continues) that the mean annual return on common stocks over the next 45 years will exceed 15%? What is the probability that the mean return will be less than 5%?

Chapter 6 Exercises

6.5 An SRS of 100 incoming freshmen was taken to look at their college anxiety level. The mean score of the sample was 83.5 (out of 100). Assuming a standard deviation of 4, give a 95% confidence interval for μ, the average anxiety level among all freshmen.

6.7 You are planning a survey of starting salaries for recent marketing majors. In 2005, the average starting salary was reported to be $37,832. Assuming the standard deviation for this study is $10,500, what sample size do you need to have a margin of error equal to $900 with 95% confidence?

6.17 For many important processes that occur in the body, direct measurement of characteristics is not possible. In many cases, however, we can measure a *biomarker*, a biochemical substance that is relatively easy to measure and is associated with the process of interest. Bone turnover is the net effect of two processes: the breaking down of old bone, called resorption, and the building of new bone, called formation. One biochemical measure of bone resorption is tartrate resistant acid phosphatase (TRAP), which can be measured in blood. In a study of bone turnover in young women, serum TRAP was measured in 31 subjects. The units are units per liter (U/l). The mean was 13.2 U/l. Assume that the standard deviation is known to be 6.5 U/l. Give the margin of error and find a 95% confidence interval for the mean for young women represented by this sample.

6.29 A new bone study is being planned that will measure the biomarker TRAP described in Exercise 6.17. Using the value of σ given there, 6.5 U/l, find the sample size required to provide an estimate of the mean TRAP with a margin of error of 2.0 U/l for 95% confidence.

6.43 You will perform a significance test of H_0: $\mu = 25$ based on an SRS of $n = 25$. Assume $\sigma = 5$.

(a) If $\bar{x} = 27$, what is the test statistic z?

(b) What is the P-value if H_A: $\mu > 25$?

(c) What is the p-value if H_A: $\mu \neq 25$?

6.57 A test of the null hypothesis H_0: $\mu = \mu_0$ gives test statistic $z = -1.73$.

(a) What is the p-value if the alternative is H_A: $\mu > \mu_0$?

(b) What is the p-value if the alternative is H_A: $\mu < \mu_0$?

(c) What is the p-value if the alternative is H_A: $\mu \neq \mu_0$?

6.69 The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. The mean score for U.S. college students is about 115, and the standard deviation is about 30. A teacher who suspects that older students have better attitudes toward school gives the SSHA to 25 students who are at least 30 years of age. Their mean score is $\bar{x} = 132.2$.

(a) Assuming that $\sigma = 30$ for the population of older students, carry out a test of H_0: $\mu = 115$ and H_A: $\mu > 115$. Report the *p*-value of your test, and state your conclusion clearly.

(b) Your test in (a) required two important assumptions in addition to the assumption that the value of σ is known. What are they? Which of these assumptions is most important to the validity of your conclusion in (a)?

6.71 Refer to Exercise 6.26. In addition to the computer computing mpg, the driver also recorded the mpg by dividing the miles driven by the number of gallons at teach fill-up. The following data are the differences between the computer's and the driver's calculations for that random sample of 20 records. The driver wants to determine if these calculations are different. Assume the standard deviation of a difference to be $\sigma = 3.0$.

5.0	6.5	–0.6	1.7	3.7	4.5	8.0	2.2	4.9	3.0
4.4	0.1	3.0	1.1	1.1	5.0	2.1	3.7	–0.6	–4.2

(a) State the appropriate H_0 and H_A to test this suspicion.

(b) Carry out the test. Give the *p*-value, and then interpret the result in plain language.

6.95 Every user of statistics should understand the distinction between statistical significance and practical importance. A sufficiently large sample will declare very small effects statistically significant. Let us suppose that SAT Mathematics (SATM) scores in the absence of coaching vary Normally with mean $\mu = 505$ and $\sigma = 100$. Suppose further that coaching may change μ but does not change σ. An increase in the SATM from 505 to 508 is of no importance in seeking admission to college, but this unimportant change can be statistically very significant. To see this, calculate the *p*-value for the test of H_0: $\mu = 505$ against H_A: $\mu > 505$ in each of the following situations:

(a) A coaching service coaches 100 students; their SATM scores average $\bar{x} = 508$.

(b) By the next year, this service has coached 1000 students; their SATM scores average $\bar{x} = 508$.

(c) An advertising campaign brings the number of students coached to 10,000; their SATM scores average $\bar{x} = 508$.

6.113 Example 6.16 gives a test of a hypothesis about the SAT scores of California high school students based on an SRS of 500 students. The hypotheses are H_0: $\mu = 450$ and H_A: $\mu > 450$. Assume that the population standard deviation is $\sigma = 100$. The test rejects H_0 at the 1% level of significance when $z \geq 2.326$, where

$$z = \frac{\bar{x} - 450}{100/\sqrt{500}}$$

Is this test sufficiently sensitive to usually detect an increase of 10 points in the population mean SAT score? Answer this question by calculating the power of the test against the alternative $\mu = 460$.

Chapter 7 Exercises

7.3 You randomly choose 15 unfurnished one-bedroom apartments from a large number of advertisements in you local newspaper. You calculate that their mean monthly rent of $570 and their standard deviation is $105. Construct a 95% confidence interval for the mean monthly rent of all advertised one-bedroom apartments.

7.5 A test of a null hypothesis versus a two-sided alternative gives $t = 2.35$.

(a) The sample size is 15. Is the test result significant at the 5% level?

(b) The sample size is 6. Is the test result significant at the 5% level?

7.25 A study of 584 longleaf pine trees in the Wade Tract in Thomas County, Georgia, is described in Example 6.1. For each tree in the tract, the researchers measured the diameter at breast height (DBH). This is the diameter of the three at 4.5 feet and the units are centimeters (cm). Only trees with DBH greater than 1.5 cm were sampled. Here are the diameters of a random sample of 40 of these trees:

10.5	13.3	26.0	18.3	52.2	9.2	26.1	17.6	40.5	31.8
47.2	11.4	2.7	69.3	44.4	16.9	35.7	5.4	44.2	2.2
4.3	7.8	38.1	2.2	11.4	51.5	4.9	39.7	32.6	51.8
43.6	2.3	44.6	31.5	40.3	22.3	43.3	37.5	29.1	27.9

(a) Use a histogram or stemplot and a boxplot to examine the distribution of DBHs. Include a Normal quantile plots if you have the necessary software. Write a careful description of the distribution.

(b) Is it appropriate to use the methods of this section to find a 95% confidence interval for the mean DBH of all trees in the Wade Tract? Explain why or why not.

(c) Report the mean and margin of error and the confidence interval.

7.29 Children in a psychology study were asked to solve some puzzles and were then given feedback on their performance. Then they were asked to rate how luck played a role in determining their scores. This variable was recorded on a 1 to 10 scale with 1 corresponding to very lucky and 10 corresponding to very unlucky. Here are the scores for 60 children:

1	10	1	10	1	1	10	5	1	1	8	1	10	2	1
9	5	2	1	8	10	5	9	10	10	9	6	10	1	5

1	9	2	1	7	10	9	5	10	10	10	1	8	1	6
10	1	6	10	10	8	10	3	10	8	1	8	10	4	2

(a) Use graphical methods to display the distribution. Describe any unusual characteristics. Do you think that these would lead you to hesitate before using the Normality-based methods of this section?
(b) Give a 95% confidence interval for the mean luck score.

7.33 Nonexercise activity thermogenesis (NEAT) provides a partial explanation for the results you found in Exercise 7.32. NEAT is energy burned by fidgeting, maintenance of posture, spontaneous muscle contraction, and other activities of daily living. In the study of Exercise 7.32, the 16 subjects increased their NEAT by 328 calories per day, on average, in response to the additional food intake. The standard deviation was 256.

(a) Test the null hypothesis that there was no change in NEAT versus the two-sided alternative. Summarize the results of the test and give your conclusion.
(b) Find a 95% confidence interval for the change in NEAT. Discuss the additional information provided by the confidence interval that is not evident from the results of the significance test.

7.35 Refer to Exercise 7.24. In addition to the computer calculating mpg, the driver also recorded the mpg by dividing the miles driven by the amount of gallons at fill-up. The driver wants to determine if these calculations are different.

Fill-up	1	2	3	4	5	6	7	8	9	10
Computer	41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
Driver	36.5	44.2	37.2	35.6	30.5	40.5	40.0	41.0	42.8	39.2
Fill-up	11	12	13	14	15	16	17	18	19	20
Computer	43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3
Driver	38.8	44.5	45.4	45.3	45.7	34.2	35.2	39.8	44.9	47.5

(a) State the appropriate H_0 and H_A.
(b) Carry out the test. Give the p-value, and then interpret the result.

7.49 Use the sign test to assess whether the computer calculates a higher mpg than the driver in Exercise 7.35. State the hypotheses, give the p-value using the binomial table (Table C), and report your conclusion.

7.57 Assume $\bar{x}_1 = 100,\ \bar{x}_2 = 120,\ s_1 = 10,\ s_2 = 12,\ n_1 = 10,\ n_2 = 10.$ Find a 95% confidence interval for the difference in the corresponding values of μ using the second approximation for degrees of freedom. Would you reject the null hypothesis that the population means are equal in favor of the two-sided alternate at significance level 0.05? Explain.

7.61 A recent study at Baylor University investigated the lipid levels in a cohort of sedentary university students. A total of 108 students volunteered for the study and met the eligibility criteria. The following table summarizes the blood lipid levels, in milligrams per deciliter (mg/dl), of the participants broken down by gender:

	Females (n = 71)		**Males (n = 37)**	
	$\bar{x}$	s	$\bar{x}$	s
Total Cholesterol	173.70	34.79	171.81	33.24
LDL	96.38	29.78	109.44	31.05
HDL	61.62	13.75	46.47	7.94

(a) Is it appropriate to use the two-sample t procedures that we studied in this section to analyze these data for gender differences? Give reasons for your answer.
(b) Describe the appropriate null and alternative hypotheses for comparing male and females total cholesterol levels.
(c) Carry out the significance test. Report the test statistic with the degrees of freedom and the p-value. Write a short summary of your conclusion.
(d) Find a 95% confidence interval for the difference between the two means. Compare the information given by the interval with the information given by the test.

7.83 A market research firm supplies manufacturers with estimates of the retail sales of their products form samples of retail stores. Marketing managers are prone to look at the estimate and ignore sampling error. Suppose that an SRS of 70 stores this month shows mean sales of 53 units of a small appliance, with standard deviation 15 units. During the same month last year, an SRS of 55 stores gave mean sales of 50 units, with standard deviation 18 units. An increase from 50 to 53 is 6%. The marketing manager is happy because sales are up 6%.

(a) Use the two-sample t procedure to give a 95% confidence interval for the difference in mean number of units sold at all retail stores.
(b) Explain in language that the marketing manager can understand why he cannot be certain that sales rose by 6%, and that in fact sales may even have dropped.

7.99 Compare the standard deviations of total cholesterol in Exercise 7.61. Give the test statistic, the degrees of freedom, and the p-value. Write a short summary of your

analysis, including comments on the assumptions of the test.

Chapter 8 Exercises

8.1 In a 2004 survey of 1200 undergraduate students throughout the United States, 89% of the respondents said they owned a cell phone. For 90% confidence, what is the margin of error?

8.3 A 1993 nationwide survey by the National Center for Education Statistics reports that 72% of all undergraduates work while enrolled in school. You decide to test whether this percent is different at your university. In your random sample of 100 students, 77 said they were currently working.

(a) Give the null and alternative hypotheses.
(b) Carry out the significance test. Report the test statistic and *p*-value.
(c) Does is appear that the percent of students working at your university is different at the $\alpha = 0.05$ level?

8.5 Refer to Example 8.6. Suppose the university was interested in a 90% confidence interval with margin of error 0.03. Would the required sample size be smaller or larger than 1068 students? Verify this by performing the calculation.

8.11 Gambling is an issue of great concern to those involved in Intercollegiate athletics. Because of this, the National Collegiate Athletic Association (NCAA) surveyed student-athletes concerning their gambling-related behaviors. There were 5594 Division I male athletes in the survey. Of these, 3547 reported participation in some gambling behavior. This included playing cards, betting on games of skill, buying lottery tickets, and betting on sports.

Find the sample proportion and the large-sample margin of error for 95% confidence. Explain in simple terms the 95%.

8.15 The Pew Poll of n = 1048 U.S. drivers found that 38% of the respondents "shouted, cursed, or made gestures to other drivers" in the last year.

Construct a 95% confidence interval for the true proportion of U.S. drivers who did these actions in the last year.

8.29 The South African mathematician John Kerrich, while a prisoner of war during World War II, tossed a coin 10,000 times and obtained 5067 heads.

(a) Is this significant evidence at the 5% level that the probability that Kerrich's coin comes up heads is not 0.5? Use a sketch of the standard Normal distribution to illustrate the *p*-value.
(b) Use a 95% confidence interval to find the range of probabilities of heads that would not be rejected at the 5% level.

8.31 Suppose after reviewing the results of the survey in Exercise 8.30, you proceeded with preliminary development of the product. Now you are at the stage where you need to decide whether or not to make a major investment to produce and market it. You will use another random sample of your customers but now you want the margin of error to be smaller. What sample size would you use if you wanted the 95% margin of error to be 0.075 or less?

8.35 A study was designed to compare two energy drink commercials. Each participant was shown the commercials in random order and asked to select the better one. Commercial A was selected by 45 out of 100 women and 80 out of 140 men. Give an estimate of the difference in gender proportions that favored Commercial A. Also construct a large-sample 95% confidence interval for this difference.

8.41 In Exercise 8.4, you were asked to compare the 2004 proportion of cell phone owners (89%) with the 2003 estimate (83%). It would be more appropriate to compare these two proportions using the methods of this section. Given that the sample size of each SRS is 1200 students, compare these to years with a significance test, and give an estimate of the difference in proportions of undergraduate cell phone owners with a 95% margin of error.

8.49 A 2005 survey of Internet users reported that 22% downloaded music onto their computers. The filing of lawsuits by the recording industry may be a reason why this percent has decreased from the estimate of 29% from a survey taken two years before. Assume that the sample sizes are both 1421. Using a significance test, evaluate whether or not there has been a change in the percent fo Internet users who download music. Provide all the details for the test and summarize your conclusion. Also report a 95% confidence interval for the difference in proportions and explain what information is provided in the interval that is not in the significance test results.

Chapter 9 Exercises

9.5 M&M Mars Company has varied the mix of colors for M&M's Milk Chocolate Candies over the years. These changes in color blends are the result of consumer preference tests. Most recently, the color distribution is reported to be 13% brown, 14% yellow, 13% red, 20% orange, 24% blue, and 16% green. You open up a 14-ounce bag of M&M's and find 61 brown, 59 yellow, 49 red, 77 orange, 141 blue, and 88 green. Use a goodness of fit test to examine how well this bag fits the percents stated by the M&M Mars company.

9.11 Cocaine addiction is difficult to overcome. Addicts have been reported to have a significant depletion of stimulating neurotransmitters and thus continue to take cocaine to avoid feelings of depression and anxiety. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium and a placebo. (Lithium is a standard drug to treat cocaine addiction. A placebo is a substance containing no medication, used so that the effect of being in the study but not taking any drug can be seen.) One-third of the subjects, chosen at random, received each treatment. Following are the results:

	Cocaine relapse?	
Treatment	**Yes**	**No**
Desipramine	10	14
Lithium	18	6
Placebo	20	4

(a) Compare the effectiveness of the three treatments in preventing relapse using percents and a bar graph. Write a brief summary.
(b) Can we comfortably use the chi-square test to test the null hypothesis that there is no difference between treatments? Explain.
(c) Perform the significance test and summarize the results.

9.17 As part of the 1999 College Alcohol Study, students who drank alcohol in the last year were asked if drinking ever resulted in missing a class. The data are given in the following table:

	Drinking Status		
Missed Class	**Nonbinger**	**Occasional Binger**	**Frequent Binger**
No	4617	2047	1176
Yes	446	915	1959

(a) Summarize the results of this table graphically and numerically.
(b) What is the marginal distribution of drinking status? Display the results graphically.

(c) Compute the relative risk of missing a class for occasional bingers versus nonbingers and for frequent bingers versus nonbingers. Summarize these results.
(d) Perform the chi-square test for this two-way table. Give the test statistic, degrees of freedom, the *p*-value, an your conclusion.

9.19 The ads in the study described in the precious exercise were also classified according to the age group of the intended readership. Here is a summary of the data:

Magazine readership age group		
Model dress	**Young adult**	**Mature adult**
Not sexual	72.3%	76.1%
Sexual	27.2%	23.9%
Number of ads	1006	503

Using parts (a) and (b) in the previous exercise as a guide, analyze these data and write a report summarizing your work.

9.25 *E. jugularis* is a type of hummingbird that lives in the forest preserves of the Caribbean island of Santa Lucia. The males and the females of this species have bills that are shaped somewhat differently. Researchers who study these birds thought that the bill shape might be related to the shape of the flowers that the visit for food. The researchers observed 49 females and 21 males. Of the females, 20 visited the flowers of *H. bihai*, while none of the males visited these flowers. Display the data in a two-way table and perform the chi-square test. Summarize the results and five a brief statement of your conclusion. Your two-way table has a count of zero in one cell. Does this invalidate your significance test? Explain why or why not.

9.31 The study of shoppers in secondhand stores cited in the previous exercise also compared the income distribution of shoppers in the two stores. Hers is the two-way table of counts:

Income	**City 1**	**City 2**
Under $10,000	70	62
$10,000 to $19,999	52	63
$20,000 to $24,999	69	50
$25,000 to $34,999	22	19
$35,000 or more	28	24

Verify that the χ^2 statistic for this table is $\chi^2 = 3.955$. Give the degrees of freedom and the *p*-value. Is there good evidence that the customers at the two stores have different income distributions?

9.35 In one part of the study described in Exercise 9.34, students were asked to respond to some questions regarding their interests and attitudes. Some of these questions form a scale called PEOPLE that measures altruism, or an interest in the welfare of others. Each student was classified as low, medium, or high on this scale. Is there an association between PEOPLE score and field of study? Here are the data:

	PEOPLE score		
Field of Study	**Low**	**Medium**	**High**
Agriculture	5	27	35
Child Dev. and Family Studies	1	32	54
Engineering	12	129	94
Liberal arts and education	7	77	129
Management	3	44	28
Science	7	29	24
Technology	2	62	64

Analyze the data and summarize your results. Are there some fields of study that have very large or very small proportions of students in the high-PEOPLE category?

9.41 The 2005 National Survey of Student Engagement reported on the use of campus services during the first year of college. In terms of academic assistance (for example tutoring, writing lab), 43% never used the services, 35% sometimes used the services,, 15% often used the services, and 7% very often used the services. You decide to see if your large university has this same distribution. You survey first-year students and obtain the counts 79, 83, 36, and 12 respectively. Use a goodness of fit test to examine how well your university reflects the national average.

Chapter 10 Exercises

10.5 The National Science Foundation collects data on the research and development spending by universities and colleges in the United States. Here are the data for the years 1999 to 2001 (using 1996 dollars):

Year	1999	2000	2001
Spending (billions of dollars)	26.4	28.0	29.7

Do the following by hand or with a calculator and verify your results with a software package.

(a) Make a scatterplot that shows the increase in research and development spending over time. Does the pattern suggest that the spending is increasing linearly over time?

(b) Find the equation of the least-squares regression line for prediction spending from year. Add this line to your scatterplot.

(c) For each of the three years, find the residual. Use these residuals to calculate the standard error s.

(d) Write the regression model for this setting What are your estimates of the unknown parameters in this model?

(e) Compute a 95% confidence interval for the slope and summarize what this interval tells you about the increase in spending over time.

10.9 For each of the settings below, test the null hypothesis that the slope is zero versus the two-sided alternate.

(a) $n = 25,\ \hat{y} = 1.3 + 12.10x,$ and $\mathrm{SE}_{b_1} = 6.31$

(b) $n = 25,\ \hat{y} = 13.0 + 6.10x,$ and $\mathrm{SE}_{b_1} = 6.31$

10.11 Refer to Exercise 10.10 and Table 10.10.

(a) Construct a 95% confidence interval for the slope. What does this interval tell you about the percent increase in tuition between 2000 and 2005?

(b) The tuition at Stat U was $5000 in 2000. What is the predicted tuition in 2005?

(c) Find a 95% prediction interval for the 2005 tuition at Stat U and summarize the results.

Table 10.1 In-state tuition and fees (in dollars) for 32 public universities					
University	Year 2000	Year 2005	University	Year 2000	Year 2005
Penn State	7018	11508	Purdue	3872	6458
Pittsburgh	7002	11436	Cal-San Diego	3848	6685
Michigan	6926	9798	Cal-Santa Barbara	3832	6997
Rutgers	6333	9221	Oregon	3819	5613
Michigan State	5432	8108	Wisconsin	3791	6284
Maryland	5136	7821	Washington	3761	5610
Illinois	4994	8634	UCLA	3698	6504
Minnesota	4877	8622	Texas	3575	6972
Missouri	4726	7415	Nebraska	3450	5540
Buffalo	4715	6068	Iowa	3204	5612
Indiana	4405	7112	Colorado	3188	5372
Ohio State	4383	8082	Iowa State	3132	5634
Virginia	4335	7370	North Carolina	2768	4613
Cal-Davis	4072	7457	Kansas	2725	5413
Cal-Berkeley	4047	6512	Arizona	2348	4498
Cal-Irvine	3970	6770	Florida	2256	3094

10.17 Consider the data in Table 10.3 and the relationship between IBI and the percent of watershed area that was forest. The relationship between these two variables is almost significant at the .05 level. In this exercise you will demonstrate the potential effect of an outlier on statistical significance. Investigate what happens when you decrease the IBI to 0.0 for (1) an observation with 0% forest and (2) an observation with 100% forest.

Table 10.3 Percent forest and index of biotic integrity									
Forest	IBI	Forest	IBI	Forest	IBI	Forest	IBI	Forest	IBI
0	47	9	33	25	62	47	33	79	83
0	61	10	46	31	55	49	59	80	82
0	39	10	32	32	29	49	81	86	82
0	59	11	80	33	29	52	71	89	86
0	72	14	80	33	54	52	75	90	79
0	76	17	78	33	78	59	64	95	67
3	85	17	53	39	71	63	41	95	56
3	89	18	43	41	55	68	82	100	85
7	74	21	88	43	58	75	60	100	91
8	89	22	84	43	71	79	84		

10.23 *Storm Data* is a publication of the National Climatic Data Center that contains a listing of tornadoes, thunderstorms, floods, lightning, temperature extremes, and

other weather phenomena. Table 10.4 summarizes the annual number of tornadoes in the United States between 1953 and 2005.

(a) Make a plot of the total number of tornadoes by year. Does a linear trend over the years appear reasonable?
(b) Are there any outliers or unusual patterns? Explain your answer.
(c) Run the simple linear regression and summarize the results, making sure to construct a 95% confidence interval for the average annual increase in the number of tornadoes.
(d) Obtain the residuals and plot them versus year. Is there anything unusual in the plot?
(e) Are the residuals Normal? Justify your answer.

Table 10.4 Annual number of tornadoes in the United States between 1953 and 2005

Year	Count	Year	Count	Year	Count	Year	Count
1953	421	1967	926	1981	783	1995	1235
1954	550	1968	660	1982	1046	1996	1173
1955	593	1969	608	1983	931	1997	1148
1956	504	1970	653	1984	907	1998	1449
1957	856	1971	888	1985	684	1999	1340
1958	564	1972	741	1986	764	2000	1076
1959	604	1973	1102	1987	656	2001	1213
1960	616	1974	947	1988	702	2002	934
1961	697	1975	920	1989	856	2003	1372
1962	657	1976	835	1990	1133	2004	1819
1963	464	1977	852	1991	1132	2005	1194
1964	704	1978	788	1992	1298		
1965	906	1979	852	1993	1176		
1966	585	1980	866	1994	1082		

10.24 In Exercise 7.26 we examined the distribution of C-reactive protein (CRP) in a sample of 40 children from Papua New Guinea. Serum retinol values for the same children were studied in Exercise 7.28. One important question that can be addressed with these data is whether or not infections, as indicated by CRP, cause a decrease in the measured values of retinol, low values of which indicate a vitamin A deficiency. The data are given in Table 10.5.

Table 10.5 C-reactive protein and serum retinol

CRP	RETINOL	CRP	RETINOL	CRP	RETINOL	CRP	RETINOL	CRP	RETINOL
0	1.15	30.61	0.97	22.82	0.24	5.36	1.19	0	0.83
3.9	1.36	0	0.67	0	1	0	0.94	0	1.11
5.64	0.38	73.2	0.31	0	1.13	5.66	0.34	0	1.02
8.22	0.34	0	0.99	3.49	0.31	0	0.35	9.37	0.56
0	0.35	46.7	0.52	0	1.44	59.76	0.33	20.78	0.82
5.62	0.37	0	0.7	0	0.35	12.38	0.69	7.1	1.2
3.92	1.17	0	0.88	4.81	0.34	15.74	0.69	7.89	0.87

6.81	0.97	26.41	0.36	9.57	1.9	0	1.04	5.53	0.41

(a) Examine the distribution of CRP and serum retinol. Use graphical and numerical methods.

(b) Forty percent of the CRP values are zero. Does this violate any assumptions that we need to do a regression analysis using CRP to predict serum retinol? Explain your answer.

(c) Run the regression, summarize the results, and write a short patagraph explaining your conclusions.

(d) Explain the assumptions needed for your results to be valid. Examine the data with respect to these assumptions and report your results.

10.37 We assume that our wages will increase as we gain experience and become more valuable to our employers. Wages also increase because of inflation. By examining a sample of employees at a given point in time, we can look at part of the picture. How does length of service (LOS) relate to wages? Table 10.8 gives data on the LOS in months and wages for 60 women who work in Indiana banks. Wages are yearly total income divided by the number of weeks worked. We have multiplied wages by a constant for reasons of confidentiality.

Table 10.8 Bank wages and length of service (LOS)

Wages	LOS	Wages	LOS	Wages	LOS
48.3355	94	64.1026	24	41.2088	97
49.0279	48	54.9451	222	67.9096	228
40.8817	102	43.8095	58	43.0942	27
36.5854	20	43.3455	41	40.7000	48
46.7596	60	61.9893	153	40.5748	7
59.5238	78	40.0183	16	39.6825	74
39.1304	45	50.7143	43	50.1742	204
39.2465	39	48.8400	96	54.9451	24
40.2037	20	34.3407	98	32.3822	13
38.1563	65	80.5861	150	51.7130	30
50.0905	76	33.7163	124	55.8379	95
46.9043	48	60.3792	60	54.9451	104
43.1894	61	48.8400	7	70.2786	34
60.5637	30	38.5579	22	57.2344	184
97.6801	70	39.2760	57	54.1126	156
48.5795	108	47.6564	78	39.8687	25
67.1551	61	44.6864	36	27.4725	43
38.7847	10	45.7875	83	67.9584	36
51.8926	68	65.6288	66	44.9317	60
51.8326	54	33.5775	47	51.5612	102

(a) Plot wages versus LOS. Describe the relationship. There is one woman with relatively high wages for her length of service. Circle this point and do not use it in the rest of this exercise.

(b) Find the least-squares line. Summarize the significance test for the slope. What do you conclude?
(c) State carefully what the slope tells you about the relationship between wages and length of service.
(d) Give a 95% confidence interval for the slope.

10.39 The Leaning Tower of Pisa is an architectural wonder. Engineers concerned about the tower's stability have done extensive studies of its increasing tilt. Measurements of the lean of the tower over time provide much useful information. The following table gives measurements for the years 1975 to 1987. The variable "lean" represents the differences between where a point on the tower would be if the tower were straight and where it actually is. The data are coded as tenths of a millimeter in excess of 2.9 meters, so that the 1975 lean, which was 2.9642 meters, appears in the table as 642. Only the last two digits of the year were entered into the computer.

Year	75	76	77	78	79	80	81	82	83	84	85	86	87
Lean	642	644	656	667	673	688	696	698	713	717	725	742	757

(a) Plot the data. Does the trend in lean over time appear to be linear?
(b) What is the equation of the least-squares line? What percent of the variation in lean is explained by this line?
(c) Give a 99% confidence interval for th average rate of change (tenths of a millimeter per year) of the lean.

10.51 A study reported a correlation $r = 0.5$ based on a sample of size $n = 20$; another reported the same correlation based on a sample size of $n = 10$. For each, perform the test of the null hypothesis that $\rho = 0$. Describe the results and explain why the conclusions are different.

Chapter 11 Exercises

11.3 Recall Exercise 11.1. Due to missing values for some students, only 86 students were used in the multiple regression analysis. The following table contains the estimated coefficients and standard errors:

Variable	Estimate	SE
Intercept	–0.764	0.651
SAT Math	0.00156	0.0007 4
SAT Verbal	0.00164	0.0007 6
High school rank	1.4700	0.430
Bryant College placement	0.889	0.402

(a) All the estimated coefficients for the explanatory variables are positive. Is this what you would expect? Explain.
(b) What are the degrees of freedom for the model and error?
(c) Test the significance of each coefficient and state your conclusions.

11.35 Let's use regression methods to predict VO+, the measure of bone formation.
(a) Since OC is a biomarker of bone formation, we start with a simple linear regression using OC as the explanatory variable. Run the regression and summarize the results. Be sure to include an analysis of the residuals.
(b) because the processes of bone formation and bone resorption are highly related, it is possible that there is some information in the bone resorption variables that can tell us something about bone formation. Use a model with both OC and TRAP, the biomarker of bone resorption, to predict VO+. Summarize the results. IN the context of this model, it appears that TRAP is a better predictor of bone formation, VO+, than the biomarker of bone formation, OC. Is this view consistent with the pattern of relationships that you described in the previous exercise? One possible explanation is that, while all of these variables are highly related, TRAP is measured with more precision than OC.

11.51 For each of the four variables in the CHEESE data set, find the mean, median, standard deviation, and interquartile range. Display each distribution by means of a stemplot and use a Normal quantile plot to assess Normality of the data. Summarize your findings. Note that when doing regressions with these data, we do not assume that these distributions are Normal. Only the residuals from our model need to be (approximately) Normal. The careful study of each variable to be analyzed is nonetheless an important first step in any statistical analysis.

11.53 Perform a simple linear regression analysis using Taste as the response variable and Acetic as the explanatory variable. Be sure to examine the residuals carefully.

Summarize your results. Include a plot of the data with the least-squares regression line. Plot the residuals versus each of the other two chemicals. Are any patterns evident? (The concentrations of the other chemicals are lurking variables for the simple linear regression.)

11.55 Repeat the analysis of Exercise 11.53 using Taste as the response variable and Lactic as the explanatory variable.

11.57 Carry out a multiple regression using Acetic and H2S to predict Taste. Summarize the results of your analysis. Compare the statistical significance of Acetic in this model with its significance in the model with Acetic alone as a predictor (Exercise 11.53). Which model do you prefer? Give a simple explanation for the fact that Acetic alone appears to be a good predictor of Taste, but with H2S in the model, it is not.

11.59 Use the three explanatory variables Acetic, H2S, and Lactic in a multiple regression to predict Taste. Write a short summary of your results, including an examination of the residuals. Based on all of the regression analyses you have carried out on these data, which model do you prefer and why?

Chapter 12 Exercises

12.3 An experiment was run to compare three groups. The sample sizes were 25, 22, and 19, and the corresponding estimated standard deviations were 22, 20, and 18.

(a) Is it reasonable to use the assumption of equal standard deviations when we analyze these data? Give a reason for your answer.
(b) Give the values of the variances for the three groups.
(c) Find the pooled variance.
(d) What is the value of the pooled standard deviation?

12.15 A study compared 4 groups with 8 observations per group. An F statistic of 3.33 was reported.

(a) Give the degrees of freedom for this statistic and the entries from Table E that correspond to this distribution.
(b) Sketch a picture of this F distribution with the information from the table included.
(c) Based on the table information, how would you report the p-value?
(d) Can you conclude that all pairs of means are different? Explain your answer.

12.17 For each of the following situations, find the F statistic and the degrees of freedom. Then draw a sketch of the distribution under the null hypothesis and shade in the portion corresponding to the p-value. State how you would report the p-value.

(a) Compare 5 groups with 9 observations per group, MSE = 50, and MSG = 127.
(b) Compare 4 groups with 7 observations per group, SSG = 40, and SSE = 153.

12.23 The National Intramural-Recreational Sports Association (NIRSA) performed a survey to look at the value of recreational sports on college campuses. One of the questions asked each student to relate the importance of recreational sports to college satisfaction and success. Responses were on a 10-point scale with 1 indicating total lack of importance and 10 indicating very high importance. The following table summarizes these results:

Class	*n*	Mean score
Freshman	724	7.6
Sophomore	536	7.6
Junior	593	7.5
Senior	437	7.3

(a) To compare the mean scores across classes, what are the degrees of freedom for the ANOVA F statistic?
(b) The MSG =11.806. If $s_p = 2.16$, what is the F statistic?
(c) Give an approximate (from a table) or exact (from software) p-value. What do you conclude?

12.25 An experimenter was interested in investigating the effects of two stimulant drugs (labeled A and B). She divided 20 rats equally into 5 groups (placebo, Drug A low, Drug A high, Drug B low, Drug B high) and 20 minutes after injection of the drug, recorded each rat's activity level (higher score is more active). The following table summarizes the results:

Treatment	$\bar{x}$	s
Placebo	14.00	8.00
Low A	15.25	12.25
High A	15.25	12.25
Low B	16.75	6.25
High B	22.50	11.00

(a) Plot the means versus the type of treatment. Does there appear to be a difference in the activity level? Explain.
(b) Is it reasonable to assume that the variances are equal? Explain your answer, and if reasonable, compute s_p.
(c) Give the degrees of freedom for the F statistic.
(d) The F statistic is 4.35. Find the associated p-value and state your conclusions.

12.29 Does bread lose its vitamins when stored? Small loaves of bread were prepared with flour that was fortified with a fixed amount of vitamins. After baking, the vitamin C content of two loaves was measured. Another two loaves were baked at the same time, stored for one day,, and then the vitamin C content was measured. In a similar manner, two loaves were stored for three, five, and seven days before measurements were taken. The units are milligrams of vitamin C per hundred grams of flour (mg/100 g). Here are the data:

Condition	**Vitamin C (mg/100 g)**	
Immediately after baking	47.62	49.79
One day after baking	40.45	43.46
Three days after baking	21.25	22.34
Five days after baking	13.18	11.65
Seven days after baking	8.51	8.13

(a) Give a table with sample size, mean, standard deviation, and standard error for each condition.
(b) Perform a one-way ANOVA for these data. Be sure to state your hypotheses, the test statistic with degrees of freedom, and the *p*-value.
(c) Summarize the data and the means with a plot. Use the plot and the ANOVA results to write a short summary of your conclusions.

12.39 Kudzu is a plant that was imported to the United States from Japan and now covers over seven million acres in the South. The plant contains chemicals called isoflavones that have been shown to have beneficial effects on bones. One study used three groups of rats to compare a control group with rats that were fed wither a low dose or a high dose of isoflavones from kudzu. One of the outcomes examined was the bone mineral density in the femur (in grams per square centimeter). Here are the data:

Treatment	**Bone mineral density (g/cm^2)**					
Control	0.228	0.221	0.234	0.220	0.217	0.228
	0.209	0.221	0.204	0.220	0.203	0.219
	0.218	0.245	0.210			
Low dose	0.211	0.220	0.211	0.233	0.219	0.233
	0.226	0.228	0.216	0.225	0.200	0.208
	0.198	0.208	0.203			
High dose	0.250	0.237	0.217	0.206	0.247	0.228
	0.245	0.232	0.267	0.261	0.221	0.219
	0.232	0.209	0.203			

(a) Use graphical and numerical methods to describe the data.
(b) Examine the assumptions necessary for ANOVA. Summarize your findings.
(c) Use a multiple-comparisons method to compare the three groups.

12.45 Recommendations regarding how long infants in developing countries should be breast-fed are controversial. If the nutritional quality of the breast milk is inadequate because the mothers are malnourished, then there is risk in inadequate nutrition for the infant. On the other hand, the introduction of other foods carries the risk of infection from contamination. Further complicating the situation is the fact that companies that produce infant formulas and other foods benefit when these foods are consumed by large numbers of customers. One question related to this controversy concerns the

amount of energy intake for infants who have other foods introduced into the diet at different ages. Part of one study compared the energy intakes, measured in kilocalories per day (kcal/d), for infants who were breast-fed exclusively for 4, 5, or 6 months. Here are the data:

Breast-fed for	Energy intake (kcal/d)						
4 months	499 517 617	620 649 704	469 209 558	485 404 653	660 738 548	588 628	675 609
5 months	490 587 368	395 528 538	402 518 519	177 370 506	475 431	617 518	616 639
6 months	585 465	647	477	445	485	703	538

(a) Make a table giving the sample size, mean, and standard deviation for each group of infants. Is it reasonable to pool the variance?
(b) Run the analysis of variance. Report the F statistic with its degrees of freedom and *p*-value. What do you conclude?

12.47 Many studies have suggested that there is a link between exercise and healthy bones. Exercise stresses the cones and this causes them to get stronger. One study examined the effect of jumping on the bone density of growing rats. There were three treatments: a control with no jumping, a low-jump condition (the jump was 30 centimeters), and a high jump condition (the ump was 60 centimeters). After 8 weeks of 10 jumps per day, 5 days per week, the bone density of the rats (expressed in mg/cm^3) was measured. Here are the data:

Group	Bone density (mg/cm^3)									
Control	611	621	614	593	593	653	600	554	603	569
Low jump	635	605	638	594	599	632	631	588	607	596
High jump	650	622	626	626	631	622	643	674	643	650

(a) Make a table giving the sample size, mean, and standard deviation for each group of rats. Is it reasonable to pool the variances?

(b) Run the analysis of variance. Report the F statistic with its degrees of freedom and p-value. What do you conclude?

12.53 Refer to Exercise 12.25. There are two comparisons of interest to the experimenter: They are (1) Placebo versus the average of the 2 low-dose treatments; and (2) the difference between High A and Low A versus the difference between High B and Low B.

(a) Express each contrast in terms of the means (μ's) of the treatments.

(b) Give estimates with standard errors for each of the contrasts.

(c) Perform the significance tests for the contrasts. Summarize the results of your tests and your conclusions.

12.63 Refer to the price promotion study that we examined in Exercise 12.40. The explanatory variable in this study is the number of price promotions in a 10 week period, with possible values of 1, 3, 5, and 7. When using analysis of variance, we treat the explanatory variable as categorical. An alternative analysis is to use simple linear regression. Perform this analysis and summarize the results. Plot the residuals from the regression model versus the number of promotions. What do you conclude?

Chapter 13 Exercises

13.7 A recent study investigated the influence that proximity and visibility of food have on food intake. A total of 40 secretaries from the University of Illinois participated in the study. A candy dish full of individually wrapped chocolates was placed either at the desk of the participant or at a location 2 meters from the participant. The candy dish was either a clear (candy visible) or opaque (candy not visible) covered bowl. After a week, the researchers noted not only the number of candies consumed per day but also the self-reported number of candies consumed by each participant. The table summarizes the mean differences between these two values (reported minus actual).

Proximity	Clear	Opaque
Proximate	–1.2	–0.8
Less proximate	0.5	0.4

Make a plot of the means and describe the patterns that you see. Does the plot suggest an interaction between visibility and proximity?

13.9 The National Crime Victimization Survey estimates that there were over 400,000 violent crimes committed against women by their intimate partner that resulted in physical injury. An intervention study designed to increase safety behaviors of abused women compared the effectiveness of six telephone intervention sessions with a control group of abused women who received standard care. Fifteen different safety behaviors were examined. One of the variables analyzed was that total number of behaviors (out of 15) that each woman performed. Here is a summary of the means of this variable at baseline (just before the first telephone call) an at follow-up 3 and 6 months later:

Group	Baseline	3 months	6 months
Intervention	10.4	12.5	11.9
Control	9.6	9.9	10.4

(a) Find the marginal means. Are they useful for understanding the results of this study?
(b) Plot the means. Do you think there is an interaction? Describe the meaning of an interaction for this study.

13.13 Analysis of data for a 3 × 2 ANOVA with 5 observations per cell gave the *F* statistics in the following table:

Effect	***F***
A	1.53
B	3.87
AB	2.94

What can you conclude from the information given?

13.17 Refer to the Exercise 13.16. Here are the standard deviations for attitude toward brand:

	Repetitions		
Familiarity	**1**	**2**	**3**
Familiar	1.16	1.46	1.16
Unfamiliar	1.39	1.22	1.42

Find the pooled estimate of the standard deviation for these data. Use the rule for examining standard deviations in ANOVA from Chapter 12 to determine if it is reasonable to use a pooled standard deviation for the analysis of these data.

13.25 One way to repair serious wounds is to insert some material as a scaffold for the body's repair cells to use as a template for new tissue. Scaffolds made form extracellular material (ECM) are particularly promising for this purpose. Because they are made form biological material, they serve as an effective scaffold and are then resorbed. Unlike biological material that includes cells, however, they do not trigger tissue rejection reactions in the body. One study compared 6 types of scaffold material. Three of these were ECMs and the other three were made of inert materials. There were three mice used per scaffold type. The response measure was the percent of glucose phosphated isomerase (Gpi) cells in the region of the wound. A large value is good, indicating that there are many bone marrow cells sent by the body to repair the tissue. In Exercise 12.51 we analyzed the data for rats whose tissues were measured 4 weeks after the repair. The experiment included additional groups of rats who received the same types of scaffold but were measured at different times. The data in the table below are for 4 weeks and 8 weeks after the repair:

(a) Make a table giving the sample size, mean, and standard deviation for each of the material-by-time combinations. Is it reasonable to pool the variances?

Because the sample sizes in this experiment are very small, we expect a large amount of variability in the sample standard deviations. Although they vary more than we would prefer, we will proceed with the ANOVA.

(b) Make a plot of the means. Describe the main features of the plot.

(c) Run the analysis of variance. Report the F statistics with degrees of freedom and *p*-values for each of the main effects and the interaction. What do you conclude?

Material	4 weeks			6 weeks		
ECM1	55	70	70	60	65	65
ECM2	60	65	65	60	70	60
ECM3	75	70	75	70	80	70
MAT1	20	25	25	15	25	25
MAT2	5	10	5	10	5	5
MAT3	10	15	10	5	10	10

13.27 Refer to the previous exercise. Analyze the data for each time perior separately using a one-way ANOVA. Use a multiple comparisons procedure where needed. Summarize the results. (The data are reproduced below.)

Material	2 weeks			4 weeks			6 weeks		
ECM1	70	75	65	55	70	70	60	65	65
ECM2	60	65	70	60	65	65	60	70	60
ECM3	80	60	75	75	70	75	70	80	70
MAT1	50	45	50	20	25	25	15	25	25
MAT2	5	10	15	5	10	5	10	5	5
MAT3	30	25	25	10	15	10	5	10	10

13.31 One step in the manufacture of large engines requires that holes of very precise dimensions be drilled. The tools that do the drilling are regularly examined and are adjusted to ensure that the holes meet the required specifications. Part of the examination involves measurement of the diameter of the drilling tool. A team studying the variation in the sizes of the drilled holes selected this measurement procedure as a possible cause of variation in the drilled holes. They decided to use a designed

experiment as one part of this examination. Some of the data are given in Table 13.2 reproduced below. The diameters in millimeters (mm) of five tools were measured by the same operator at three times (8:00 a.m., 11:00 a.m., and 3:00 p.m.). The person taking the measurements could not tell which tool was being measured, and the measurements were taken in random order.

(a) Make a table of means and standard deviations for each of the 5×3 combinations of the two factors.

(b) Plot the means and describe how the means vary with tool and time. Note that we expect the tools to have slightly different diameters. These will be adjusted as needed. It is the process of measuring the diameters that is important.

(c) Use a two-way ANOVA to analyze these data. Report the test statistics, degrees of freedom, and p-values for the significance tests.

Tool	Time	Diameter (mm)		
1	1	25.030	25.030	25.032
1	2	25.028	25.028	25.028
1	3	25.026	25.026	25.026
2	1	25.016	25.018	25.016
2	2	25.022	25.020	25.018
2	3	25.016	25.016	25.016
3	1	25.005	25.008	25.006
3	2	25.012	25.012	25.014
3	3	25.010	25.010	25.008
4	1	25.012	25.012	25.012
4	2	25.018	25.020	25.020
4	3	25.010	25.014	25.018
5	1	24.996	24.998	24.998
5	2	25.006	25.006	25.006
5	3	25.000	25.002	24.999

13.35 A study of the question "Do left-handed people live shorter lives than right-handed people?" examined a sample of 949 death records and contacted next of kin to determine handedness. Note that there are many possible definitions of "left-handed." The researchers examined the effects of different definitions on the results of their analysis and found that their conclusions were not sensitive to the exact definition used. For the results presented here, people were defined to be right-handed if they wrote, drew, and threw a ball with the right hand. All others were defined to be left-handed. People were classified by gender (female or male), and a 2×2 ANOVA was run with the age at death as the response variable. The F statistics were 22.36 (handedness), 37.44 (gender), and 2.10 (interaction). The following marginal mean ages at death (in years) were reported: 77.39 (females), 71.32 (males), 75.00 (right-handed), and 66.03 (left-handed).

(a) For each of the F statistics given above find the degrees of freedom and an approximate P-value. Summarize the results of these tests.

Chapter 14 Exercises

14.1 If you deal one card from a standard deck, the probability that the card is a heart is 0.25. Find the odds of drawing a heart.

14. 3 A study was designed to compare two energy drink commercials. Each participant was shown two commercials, A and B, in random order and asked to select the better one. There were 100 women and 140 men who participated in the study. Commercial A was selected by 45 women and by 80 men. Find the odds of selecting Commercial A for the men. Do the same for the women.

14.5 Refer to Exercise 14.3. Find the log odds for the men and the log odds for the women.

14.7 Refer to Exercises 14.3 and 14.5. Find the logistic regression equation and the odds ratio.

14.11 Following complaints about the working conditions in some apparel factories both in the United States and abroad, a joint government and industry commission recommended in 1998 that companies that monitor and enforce proper standards be allowed to display a "No Sweat" label on their products. Does the presence of these labels influence consumer behavior?

A survey of U.S. residents aged 18 or older asked a series of questions about how likely they would be to purchase a garment under various conditions. For some conditions, it was stated that the garment had a "No Sweat" label; for other, there was no mention of such a label. On the basis of the responses, each person was classified as a "label user" or a "label nonuser." Suppose we want to examine the data for a possible gender effect. Here are the data for comparing men and women:

Gender	*n*	Number of Label users
Women	296	63
Men	251	27

(a) For each gender find the proportion of label users.
(b) Convert each of the proportions that you found in part (a) to odds.
(c) Find the log of each of the odds that you found in part (b).

14.13 Refer to Exercise 14.11. Use $x = 1$ for women and $x = 0$ for men.

(a) Find the estimates b_0 and b_1.

(b) Give the fitted logistic regression model.

(c) What is the odds ratio for men versus women?

14.21 Different kinds of companies compensate their key employees in different ways. Established companies may pay higher salaries, while new companies may offer stock options that will be valuable if the company succeeds. Do high-tech companies tend to offer stock options more often than other companies? One study looked at a random sample of 200 companies. Of these, 91 were listed in the *Directory of Public High Technology Corporations*, and 109 were not listed. Treat these two groups as SRSs of high-tech and non-high-tech companies. Seventy-three of the high-tech companies and 75 of the non-high-tech companies offered incentive stock options to key employees.

(a) What proportion of the high-tech companies offer stock options to their key employees? What are the odds?

(b) What proportion of the non-high-tech companies offer stock options to their key employees? What are the odds?

(c) Find the odds ratio using the odds for the high-tech companies in the numerator. Describe the result in a few sentences.

14.25 There is much evidence that high blood pressure is associated with increased risk of death from cardiovascular disease. A major study of this association examined 3338 men with high blood pressure and 2676 men with low blood pressure. During the period of the study, 21 men from the low-blood-pressure group and 55 in the high-blood-pressure group died from cardiovascular disease.

(a) Find the proportion of men who died from cardiovascular disease in the high-blood-pressure group. Then calculate the odds.

(b) Do the same for the low-blood-pressure group.

(c) Now calculate the odds ratio with the odds for the high-blood-pressure group in the denominator. Describe the result in words.

14.27 Refer to the study of cardiovascular disease and blood pressure in Exercise 14.25. Computer output for a logistic regression analysis of these data gives an estimated slope $b_1 = 0.7505$ with standard error $SE_{b_1} = 0.2578$.

(a) Five a 95% confidence interval for the slope.

(b) Calculate the X^2 statistic for testing the null hypothesis that the slope is zero and use Table F to find an approximate p-value.

14.35 A study of alcohol use and deaths due to bicycle accidents collected data on a large number of fatal accidents. For each of these, the individual who died was classified according to whether or not there was a positive test for alcohol and by gender. Here are the data:

Gender	*n*	X (tested positive)

Female	191	27
Male	1520	515

Use logistic regression to study the question of whether or not gender is related to alcohol use in people who are fatally injured in bicycle accidents.

Chapter 15 Exercises

15.3 Refer to Exercise 15.1. State appropriate null and alternative hypotheses for this setting and calculate the value of *W*, the test statistic.

Group A	552	448	68	243	30
Group B	329	780	560	540	240

15.5 Refer to Exercises 15.1 and 15.3. Find μ_W, σ_W, and the standardized rank sum statistic. Then give the approximate p-value using the Normal approximation. What do you conclude?

15.11 How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed. Part of the study involved burying 10 polyester strips in well-drained soil in the summer. Five of the strips, chosen at random, were dug up after 2 weeks; the other 5 were dug up after 16 weeks. Here are the breaking strengths in pounds:

2 weeks	118	126	126	120	129
16 weeks	124	98	110	140	110

(a) Make a back-to-back stemplot. Does it appear reasonable to assume that the two distributions have the same shape?
(b) Is there evidence that the breaking strengths are lower for the strips buried longer?

15.19 Refer to Exercise 15.18. Here are the scores for a random sample of 7 spas that ranked between 19 and 36:

Spa	1	2	3	4	5	6	7
Diet/Cuisine	77.3	85.7	84.2	85.3	83.7	84.6	78.5
Program/Facilities	95.7	78.0	87.2	85.3	93.6	76.0	86.3

Is food, expressed by the Diet/Cuisine score, more important than activities, expressed as the Program/Facilities score, for a top ranking? Formulate this question in terms of null and alternative hypotheses. Then compute the differences and find the value of the Wilcoxon signed rank statistic, W^+.

15.21 Refer to exercise 15.19. Find μ_{W^+}, σ_{W^+}, and the Normal approximation for the p-value for the Wilcoxon signed rank test.

15.25 Can the full moon influence behavior? A study observed at nursing home patients with dementia. The number of incidents of aggressive behavior was recorded each dat for 12 weeks. Call a day a "moon day" if it is the day of a full moon or the day before or after a full moon. Here are the average numbers of aggressive incidents for moon days and other days for each subject:

Patient	Moon days	Other days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26
6	3.67	0.11
7	4.67	0.30
8	2.67	0.40
9	6.00	1.59
10	4.33	0.60
11	3.33	0.65
12	0.67	0.69
13	1.33	1.26
14	0.33	0.23
15	2.00	0.38

The matched pairs t test (Example 7.7) gives $P < 0.000015$ and a permutation test (Example 16.14) gives $P = 0.0001$. Does the Wilcoxon signed rank test, based on ranks rather than means, agree that there is strong evidence that there are more aggressive behaviors on moon days?

15.31 Exercise 7.32 presents the data below on the weight gains (in kilograms) of adults who were fed an extra 1000 calories per day for 8 weeks.

(a) Use a rank test to test the null hypothesis that the median weight gain is 16 pounds, as theory suggests. What do you conclude?

Subject	Before	After
1	55.7	61.7
2	54.9	58.8
3	59.6	66.0
4	62.3	66.2
5	74.2	79.0
6	75.6	82.3
7	70.7	74.3
8	53.3	59.3
9	73.3	79.1
10	63.4	66.0
11	68.1	73.4
12	73.7	76.9
13	91.7	93.1
14	55.9	63.0
15	61.7	68.2
16	57.8	60.3

15.33 Many studies suggest that exercise causes bones to get stronger. One study examined the effect of jumping on the bone density of growing rats. Ten rats were assigned to each of three treatments: a 60-centimeter "high jump," a 30-centimeter "low jump," and a control group with no jumping. Here are the bone densities (in milligrams per cubic centimeter) after 8 weeks of 10 jumps per day:

Group	Bone density (mg/cm^3)									
Control	611	621	614	593	593	653	600	554	603	569
Low jump	635	605	638	594	599	632	631	588	607	596
High jump	650	622	626	626	631	622	643	674	643	650

(c) Do the Kruskal-Wallis test. Explain the distinction between thehypotheses tested by Kruskal-Wallis and ANOVA.

Chapter 16 Exercises

16.5 The distribution of carbon dioxide (CO_2) emissions in Table 1.6 is strongly skewed to the right. The United States and several other countries appear to he high outliers. Generate a bootstrap distribution for the mean of CO_2 emissions; construct a histogram and Normal quantile plot to assess Normality of the bootstrap distribution. On the basis of your work, do you expect the sampling distribution of $\bar{x}$ to be close to Normal?

16.7 The measurements of C-reactive protein in 40 children (Exercise 7.26) are very strongly skewed. We were hesitant to use *t* procedures for these data. Generate a bootstrap distribution for the mean of C-reactive protein; construct a histogram and Normal quantile plot to assess Normality of the bootstrap distribution. On the basis of your work, do you expect the sampling distribution of $\bar{x}$ to be close to Normal?

16.9 We have two ways to estimate the standard deviation of a sample mean $\bar{x}$: use the formula $s/\sqrt{n}$ for the standard error, or use the bootstrap standard error.

(b) Find the sample standard deviation s for the CO_2 emissions in Exercise 16.5 and use it to find the standard error $s/\sqrt{n}$ of the sample mean. How closely does your result agree with the bootstrap standard error from your resampling in Exercise 16.5?

16.13 Return to or create the bootstrap distribution resamples on the sample mean for the audio file lengths in Exercise 16.8. In Example 7.11, the *t* confidence interval for the average length was constructed.

(a) Inspect the bootstrap distribution. Is a bootstrap *t* confidence interval appropriate? Explain why or why not.

(b) Construct the 95% bootstrap *t* confidence interval.

(c) Compare the bootstrap results with the *t* confidence interval reported in Example 7.11.

16.25 Each year, the business magazine *Forbes* publishes a list of the world's billionaires. In 2006, the magazine found 793 billionaires. Here is the wealth, as estimated by *Forbes* and rounded to the nearest 100 million, of an SRS of 20 of these billionaires:

2.9	15.9	4.1	1.7	3.3	1.1	2.7	13.6	2.2	2.5
3.4	4.3	2.7	1.2	2.8	1.1	4.4	2.1	1.4	2.6

Suppose you are interested in "the wealth of typical billionaires." Bootstrap an appropriate statistic, inspect the bootstrap distribution, and draw conclusions based on this sample.

16.31 Consider the small random subset of the Verizon data in Exercise 16.1. Bootstrap the sample mean using 1000 resamples. The data are reproduced below:

26.47	0.00	5.32	17.30	29.78	3.67

(a) Make a histogram and Normal quantile plot. Does the bootstrap distribution appear close to Normal? Is the bias small relative to the observed sample mean?
(b) Find the 95% bootstrap *t* confidence interval.
(c) Five the 95% bootstrap percentile confidence interval and compare it with the interval in part (b).

16.45 Figure 2.7 shows a very weak relationship between returns on Treasury bills and returns on common stocks. The correlation is $r = -0.113$. We wonder if this is significantly different from 0. To find out, bootstrap the correlation. (The data are in the file ex16-045.)

(a) Describe the shape and bias of the bootstrap distribution. It appears that even simple bootstrap inference (t and percentile confidence intervals) is justified. Explain why.

16.59 Exercise 7.41 gives data on a study of the effect of a summer language institute on the ability of high school language teachers to understand spoken French. This is a matched pairs study, with scores for 20 teachers at the beginning (pretest) and end (posttest) of the institute. We conjecture that the posttest scores are higher on the average.

(a) Carry out the matched pairs t test. That is, state the hypotheses, calculate the test statistic, and give its p-value.
(b) Make a Normal quantile plot of the gains: posttest score – pretest score. The data have a number of ties and a low outlier. A permutation test can help check the *t* test result.
(c) Carry out the permutation test for the difference in means in matched pairs, using 9999 resamples. The Normal quantile plot shows that the permutation distribution is reasonably Normal, but the histogram looks a bit odd. What explains the appearance of the histogram? What is the *P*-value for the permutation test? Do your tests in here and in part (a) lead to the same practical conclusion?

16.77 Exercise 2.17 (page 97) describes a study that suggests that the "pain" caused by social rejection really is pain, in the sense that it causes activity in brain areas known to be activated by physical pain. Here are data for 13 subjects on degree of social distress and extent of brain activity.

Subject	Social distress	Brain activity	Subject	Social distress	Brain activity
1	1.26	–0.055	8	2.18	0.025
2	1.85	–0.040	9	2.58	0.027
3	1.10	–0.026	10	2.75	0.033
4	2.50	–0.017	11	2.75	0.064
5	2.17	–0.017	12	3.33	0.077
6	2.67	0.017	13	3.65	0.124
7	2.01	0.021			

Make a scatterplot of brain activity against social distress. There is a positive linear association with correlation r = 0.878. Is this correlation significantly greater than 0? Use a permutation test.

16.85 The researchers in the study described in the Exercise 16.84 expected higher word counts in magazines aimed at people with high education level. Do a permutation test to see if the data support this expectation. State hypotheses, give a p-value, and state your conclusions. How do your conclusions here relate to those from Exercise 16.84?

Education level	Word count								
High	205 80	203 208	229 89	208 49	146 93	230 46	215 34	153 39	205 88
Medium	191 94	219 206	205 197	5768	105 44	109 203	8213 9	8872	3967

Chapter 17 Exercises

17.5 A sandwich shop owner takes a daily sample of 6 consecutive sandwich orders at random times during the lunch rush and records the time it takes to complete each order. Past experience indicates that the process mean should be $\mu = 168$ seconds and the process standard deviation should be $\sigma = 30$ seconds. Calculate the center line and control limits for an $\bar{x}$ control chart.

17.13 A meat-packaging company produces 1-pound packages of ground beef by having a machine slice a long circular cylinder of ground beef as it passes through the machine. The timing between consecutive cuts will alter the weight of each section. Table 17.3, reproduced below, gives the weight of 3 consecutive sections of ground beef taken each hour over two 10-hour days. Past experience indicates that the process mean is 1.03 and the weight varies with $\sigma = 0.02$ lb.

Sample	Weight (pounds)			$\bar{x}$	s
1	0.999	1.071	1.019	1.030	0.0373
2	1.030	1.057	1.040	1.043	0.0137
3	1.024	1.020	1.041	1.028	0.0108
4	1.005	1.026	1.039	1.023	0.0172
5	1.031	0.995	1.005	1.010	0.0185
6	1.020	1.009	1.059	1.029	0.0263
7	1.019	1.048	1.050	1.039	0.0176
8	1.005	1.003	1.047	1.018	0.0247
9	1.019	1.034	1.051	1.035	0.0159
10	1.045	1.060	1.041	1.049	0.0098
11	1.007	1.046	1.014	1.022	0.0207
12	1.058	1.038	1.057	1.051	0.0112
13	1.006	1.056	1.056	1.039	0.0289
14	1.036	1.026	1.028	1.030	0.0056
15	1.044	0.986	1.058	1.029	0.0382
16	1.019	1.003	1.057	1.026	0.0279
17	1.023	0.998	1.054	1.025	0.0281
18	0.992	1.000	1.067	1.020	0.0414
19	1.029	1.064	0.995	1.029	0.0344
20	1.008	1.040	1.021	1.023	0.0159

(a) Calculate the center line and control limits for an $\bar{x}$ chart.
(b) What are the center line and control limits for an s chart for this process?
(c) Create the $\bar{x}$ and s chards for these 20 consecutive samples.
(d) Does the process appear to be in control? Explain.

17.15 A pharmaceutical manufacturer forms tablets by compressing a granular material that contains the active ingredient and various fillers. The hardness of a sample from each lot of tables is measured in order to control the compression process. The process has been operating in control with mean at the target value $\mu = 11.5$ and estimated standard deviation $\sigma = 0.2$. Table 17.4 gives three sets of data, each representing $\bar{x}$ for 20 successive samples of n = 4 tablets. One set of data remains in control at the target value. In a second set, the process mean μ shifts suddenly to a new value. In a third, the process mean drifts gradually.

(a) What are the center line and control limits for an $\bar{x}$ chart for this process?

(b) Draw a separate $\bar{x}$ chart for each of the three data sets. Mark any points that are beyond the control limits.

(c) Based on your work in (b) and the appearance of the control charts, which set of data comes from a process that is in control? In which case does the process mean shift suddenly, and at about which sample do you think that the mean changed? Finally, in which case does the mean drift gradually?

Sample	Data A	Data B	Data C
1	11.602	11.627	11.495
2	11.547	11.613	11.475
3	11.312	11.493	11.465
4	11.449	11.602	11.497
5	11.401	11.360	11.573
6	11.608	11.374	11.563
7	11.471	11.592	11.321
8	11.453	11.458	11.533
9	11.446	11.552	11.486
10	11.522	11.463	11.502
11	11.664	11.383	11.534
12	11.823	11.715	11.624
13	11.629	11.485	11.629
14	11.602	11.509	11.575
15	11.756	11.429	11.730
16	11.707	11.477	11.680
17	11.612	11.570	11.729
18	11.628	11.623	11.704
19	11.603	11.472	12.052
20	11.816	11.531	11.905

17.19 Figure 17.10 reproduces a data sheet from the floor of a factory that makes electrical meters. The sheet shows measurements of the distance between two mounting holes for 18 samples of size 5. The heading informs us that the measurements are in multiples of 0.0001 inch above 0.6000 inch. That is, the first measurement, 44, stands for 0.6044 inch. All the measurements end in 4. Although we don't know why this is true,

it is clear that in effect the measurements were made to the nearest 0.001 inch, not to the nearest 0.0001 inch.

Calculate $\bar{x}$ and s for the first two samples. The data file *ex17_19* contains $\bar{x}$ and s for all 18 samples. Based on long experience with this process, you are keeping control charts based on $\mu = 43$ and $\sigma = 12.74$. Make s and $\bar{x}$ charts for the data in Figure 17.10 and describe the state of the process.

17.21 An $\bar{x}$ chart plots the means of samples of size 4 against center line CL = 700 and control limits LCL = 685 and UCL = 715. The process has been in control.

(a) What are the process mean and standard deviation?

(b) The process is disrupted in a way that changes the mean to $\mu = 690$. What is the probability that the first sample after the disruption gives a point beyond the control limits of the $\bar{x}$ chart?

(c) The process is disrupted in a way that changes the mean to $\mu = 690$ and the standard deviation to $\sigma = 15$. What is the probability that the first sample after the disruption gives a point beyond the control limits of the $\bar{x}$ chart?

17.31 The $\bar{x}$ and s control charts for the mesh-tensioning example (Figures 17.4 and 17.7) were based on $\mu = 275$ mV and $\sigma = 43$ mV. Table 17.1 gives the 20 most recent samples from this process.

(a) Estimate the process μ and σ based on these 20 samples.

(b) Your calculations suggest that the process σ may now be less than 43 mV. Explain why the s chart in Figure 17.7 (page 17-15) suggests the same conclusion. (If this pattern continues, we would eventually update the value of σ used for control limits.)

17.35 Do the losses on the 120 individual patients in Table 17.7 appear to come from a single Normal distribution? Make a Normal quantile plot and discuss what it shows. Are the natural tolerances you found in the Exercise 17.34 trustworthy?

17.37 The center of the specification for mesh tension is 250 mV, but the center of our process is 275 mV. We can improve capability by adjusting the process to have center 250 mV. This is an easy adjustment that does not change the process variation. What percent of monitors now meet the new specifications? (From Exercise 17.36, the specifications are 150 to 350 mV; the standard deviation is 38.4 mV.)

17.41 The record sheet in Figure 17.10 gives specifications as 0.6054 ± 0.0010 inch. That's 54 ± 10 as the data are coded on the record. Assuming that the distance varies Normally from meter to meter, about what percent of meters meet the specifications?

17.43 Make a Normal quantile plot of the 85 distances in data file *ex17_19* that remain after removing sample 5. How does the plot reflect the limited precision of the measurements (all of which end in 4)? Is there any departure from Normality that would lead you to discard your conclusions from Exercise 17.39?

17.53 Table 17.1 gives 20 process control samples of the mesh tension of computer monitors. In Example 17.13, we estimated from these samples that $\hat{\mu} = \bar{\bar{x}} = 275.065$ mV and $\hat{\sigma} = s = 38.38$ mV.

(a) The original specifications for mesh tension were LSL = 100 mV and USL = 400 mV. Estimate C_p and C_{pk} for this process.

(b) A major customer tightened the specifications to LSL = 150 mV and USL = 350 mV. Now what are C_p and C_{pk}?

17.71 An egg farm wants to monitor the effects of some new handling procedures on the percent of eggs arriving at the packaging center with cracked or broken shells. In th past, roughly 2% of the eggs were damaged. A machine will allow the farm to inspect 500 eggs per hour. What are the initial center line and control limits for a chart of the hourly percent of damaged eggs?

17.77 Because the manufacturing quality in the Exercise 17.76 is so high, the process of writing up orders is the major source of quality problems: the defect rate there is 8000 per million opportunities. The manufacturer processes about 500 orders per month.

(a) What is $\bar{p}$ for the order-writing process? How many defective orders do you expect to see in a month?

(b) What are the center line and control limits for a p chart for plotting monthly proportions of defective orders? What is the smallest number of bad orders that will result in a point above the upper control limit?

17.83 You have just installed a new system that uses an interferometer to measure the thickness of polystyrene film. To control the thickness, you plan to measure 3 film specimens every 10 minutes and keep $\bar{x}$ and s charts. To establish control, you measure 22 samples of 3 films each at 10-minute intervals. Table 17.12 gives $\bar{x}$ and s for these samples. The units are millimeters $\times 10^{-4}$. Calculate control limits for s, make an s chart, and comment on control of short-term process variation.