MINITAB Manual

Michael Evans
University of Toronto

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: 0-7167-2994-6

Contents

III. Exercises 239

Preface

This manual serves as an introduction to the statistical software package Minitab. It is targeted at students who are taking an introductory statistics course. The manual covers the computational topics typically needed in such a course.

Minitab is easy to learn and use. The time students need to spend to learn Minitab, and that instructors need to allocate to teach it, is relatively small. Also Minitab serves as a perfectly adequate tool for many of the statistical computational problems students will encounter throughout their undergraduate education.

This manual can be used with either Minitab Version 15, Minitab Student Version 14, Minitab Version 14 or Minitab Version 13 running under Windows. The text is based on Minitab Version 15. The core of the manual is a discussion of the menu commands while not neglecting to refer to the session commands, as these are needed for certain problems. The material on session commands is always at the end of each section and can be skipped if the reader will definitely not be using them. We have provided some exercises for each chapter.

The manual is divided into two parts. Part I is an introduction that provides the necessary details to start using Minitab and, in particular, explains how to use worksheets. We recommend reading Part I before starting to use Minitab. Overall, the introductory Part I serves as a reference for most of the nonstatistical commands in Minitab and is basically concerned with Data Management.

Part II introduces the statistical commands. The sequence of chapters follows the organization of a typical introductory statistics course. The Minitab commands relevant to doing typical problems encountered in such a course are introduced and their use illustrated. Each chapter concludes with a set of exercises. These are specifically designed to ensure that the relevant Minitab material has been understood.

This manual does not attempt a complete coverage of Minitab. Rather, we introduce and discuss those concepts in Minitab that we feel are most relevant for a student studying introductory statistics. While the manual's primary goal is to teach Minitab, generally we want to help develop strong data analytic skills.

For further information on Minitab software, contact:

Minitab Inc.
3081 Enterprise Drive
State College, PA 16801 USA
ph: 814.328.3280
fax: 814.238.4383
email: Info@minitab.com
URL: http://www.minitab.com

Part I

Minitab for Data Management

New Minitab commands discussed in this part

Calc ▶ Calculator
Calc ▶ Column Statistics
Calc ▶ Make Patterned Data
Calc ▶ Row Statistics
Edit ▶ Copy Cells
Edit ▶ Cut Cells
Edit ▶ Paste Cells
Edit ▶ Select All Cells
Edit ▶ Undo Cut
Edit ▶ Undo Paste
Editor ▶ Enable Commands
Editor ▶ Insert Cells
Editor ▶ Insert Columns
Editor ▶ Insert Rows
Editor ▶ Output Editable
File ▶ Exit
File ▶ New
File ▶ Other Files ▶ Export Special Text
File ▶ Open Worksheet
File ▶ Other Files ▶ Import Special Text
File ▶ Print Session Window
File ▶ Print Worksheet
File ▶ Save Current Worksheet
File ▶ Save Current Worksheet As
File ▶ Save Session Window As
Help
Data ▶ Copy Columns
Data ▶ Display Data
Data ▶ Erase Variables
Data ▶ Rank
Data ▶ Sort
Window ▶ Project Manager

1 Manual Overview and Conventions

Minitab is a software package for carrying out statistical, numerical, and graphical calculations. This manual does not attempt to describe all the possible implementations or the full extent of the package. We limit our discussion to those features common to the most recent versions of Minitab running under the Windows operating system. Version 15 refers to the latest version of Minitab at the time of writing this manual, but we also make reference to Versions 13 and 14 when there are differences. This manual can be used with each version.

In this manual, special statistical or Minitab concepts will be highlighted in *italic* font. You should be sure that you understand these concepts.

Primarily, we will be discussing the *menu commands* that are available in Minitab. Menu commands are accessed by clicking the left button of the mouse

on items in lists. We use a special notation for menu commands. For example,

A ▶ B ▶ C

is to be interpreted as left click the command A on the menu bar, then in the list that drops down, left click the command B, and, finally, left click C. The menu commands will be denoted in ordinary font (the actual appearance may vary slightly depending on the version of Windows you use).

There are also *session commands* and *subcommands* that are typed by the user rather than using the mouse. These will be denoted in **bold** font. Any commands that we actually type, and the output obtained, will be denoted in `typewriter` font, as will the names of any files used by Minitab, variables, constants, and worksheets.

We recommend that whenever feasible, the reader use Minitab to do the problems in your text. While many problems can be done by hand, you will save a considerable amount of time and avoid errors by learning to use Minitab effectively. We also recommend that you try out the Minitab commands as you read about them, as this will ensure full understanding.

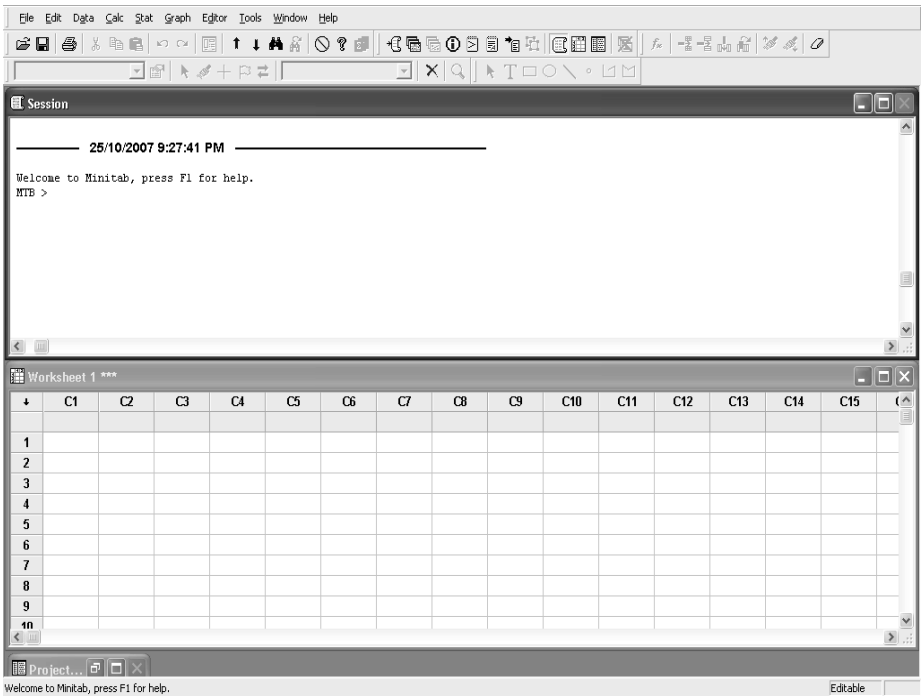
2 Accessing and Exiting Minitab

The first thing you should do is find out how to access the Minitab package. This information will come from your instructor, system personnel, or from your software documentation if you have purchased Minitab to run on your own computer.

In most cases, you will double click an icon, such as that shown in Display I.1, that corresponds to the Minitab program. Alternatively, you can use the Start button and click on Minitab in the Programs list. In this case, the program opens with a *Minitab window,* such as the one shown in Display I.2. The Minitab window is divided into two sub-windows with the upper window called the *Session window* and the lower one called the *Data window.*

Left clicking the mouse anywhere on a particular window brings that window to the foreground—i.e., makes it the *active* window—and the border at the top of the window turns dark blue. For example, clicking in the Session window will make that window active. Alternatively, you can use the command Window ▶ Session in the *menu bar* at the top of the Minitab window to make this window active.

Display I.1: Minitab icon.



Display I.2: Minitab window.

You may not see the

```
MTB >
```

prompt in the Session window, and for some things described in this manual it is important that you do so. You can ensure that this prompt always appears in your Session window by using Tools ▶ Options ▶ Session Window ▶ Submitting Commands, clicking on the Enable radio button and then clicking on OK. Without the `MTB >` prompt, you cannot type commands to be executed in the Session window.

In the session window, Minitab *commands* are typed after the

```
MTB >
```

prompt and executed when you hit the Enter or Return key. For example, the command **exit** takes you out of your Minitab session and returns you to the system prompt or operating system. Otherwise, you can access commands using the menu bar (Display I.3) that resides at the top of the Minitab window. For example, you can access the **exit** command using File ▶ Exit. In many circumstances, using the menu commands to do your analyses is easy and convenient, although there are certain circumstances where typing the session commands is necessary. You can also exit by clicking on the × symbol in the upper right-hand corner of the Minitab window. When you exit, you are prompted by Minitab in a dialog window with something like the question, "Save changes to the Project

'Untitled' before closing?" You can safely answer no to this question unless you are in fact using the Projects feature in Minitab as described in Appendix A. Later, we will discuss how to save the contents of a Data window before exiting. This is something you will commonly want to do.

Display I.3: Menu bar.

Immediately below the menu bar in the Minitab window is the *taskbar*. The taskbar consists of various icons that provide a shortcut method for carrying out various operations by clicking on them. These operations can be identified by holding the cursor over each in turn, and it is a good idea to familiarize yourself with these as they can save time. Of particular importance are the Cut, Copy, and Paste icons, which are available when a Data window is active. When the operation associated with an icon is not available, the icon is faded.

Minitab is an interactive program. By this we mean that you supply Minitab with input data, or tell it where your input data is, and then Minitab responds instantaneously to any commands you give telling it to do something with that data. You are then ready to give another command. It is also possible to run a collection of Minitab commands in a batch program; i.e., several Minitab commands are executed sequentially before the output is returned to the user. The batch version is useful when there is an extensive number of computations to be carried out. You are referred to Help on the menu bar if you want to use this feature.

3 Files Used by Minitab

Minitab can accept input from a variety of files and write output to a variety of files. Each file is distinguished by a *file name* and an *extension* that indicates the type of file it is. For example, `marks.mtw` is the name of a file that would be referred to as 'marks' (note the single quotes around the file name) within Minitab. The extension `.mtw` indicates that this is a Minitab worksheet. We describe what a worksheet is in Section I.5. This file is stored somewhere on the hard drive of a computer as a file called `marks.mtw.`

There are other files that you will want to access from outside Minitab, perhaps to print them out on a printer. In such a case, you have to give the relevant system print command together with the full path name of the file you wish to print. As various implementations of Minitab differ as to where these files are stored on the hard drive, you will have to determine this information from your instructor or documentation or systems person. For example, in Windows the full path name of the worksheet file `marks.mtw` could be

```
C:\minitabdata\marks.MTW
```

or something similar. This path name indicates that the file `marks.mtw` is stored

on the C hard drive in the directory called \minitabdata, which I created. We will discuss several different types of files in this manual.

It is generally best to name your files so that the file name reflects its contents. For example, the file name `marks` may refer to a data set composed of student marks in a number of courses.

4 Getting Help

At times, you may want more information about a command or some other aspect of Minitab than this manual provides, or you may wish to remind yourself of some detail that you have partially forgotten. Minitab contains an online manual that is very convenient. You can access this information directly by clicking on Help in the Menu bar and using the table of contents (via Help ▶ Help) or doing a search (via Search ▶ Help) of the manual for a particular concept.

From the

```
MTB >
```

prompt, you can use the **help** command for this purpose. Typing **help** followed by the name of the command of interest and hitting Enter will cause Minitab to produce a window containing relevant output. For example, asking for help on the command **help** itself via the command

```
MTB >help help
```

will give you the table of contents of the online help manual with **help** highlighted. The **help** command should be used to find out about session commands.

5 The Worksheet

The basic structural component of Minitab is the *worksheet*. Basically, the worksheet can be thought of as a big rectangular array, or matrix, of *cells* organized into rows and columns as in the Data window of Display I.2. Each cell holds one piece of data. This piece of data could be a number, i.e., *numeric data*, or it could be a sequence of characters, such as a word or an arbitrary sequence of letters and numbers, i.e., *text data*. Data often comes as numbers, such as $1.7, 2.3, \ldots$, but sometimes it comes in the form of a sequence of characters, such as black, brown, red, etc. Typically, sequences of characters are used as identifiers in classifications for some variable of interest; for example, color, gender. A piece of text data can be up to 80 characters in length in Minitab. Minitab also allows for *date data*, which is data especially formatted to indicate a date, for example, 3/4/97. We will not discuss date data.

If possible, try to avoid using text data with Minitab; i.e., make sure all the values of a variable are numbers, as dealing with text data in Minitab is more difficult. For example, denote colors by numbers rather than by names. Still, there will be applications where data comes to you as text data—for example, in

a computer file—and it is too extensive to convert to numeric data. So we will discuss how to input text data into a Minitab worksheet, but we recommend that in such cases you convert text data to numeric data, using the methods of Section C.3 in Appendix C, once it has been input.

Display I.4 provides an example of part of a worksheet. Notice that the columns are labeled C1, C2, etc., and the rows are labeled 1, 2, 3, etc. We will refer to the worksheet depicted in Display I.4 as the `marks` worksheet hereafter and will use it throughout Part I to illustrate various Minitab commands and operations.

↓	C1	C2	C3	C4	C5-T	C6
3	53546	77	83	81	f	
4	55542	63	42	56	m	
5	11223	71	82	67	f	
6	77788	87	56	*	f	
7	44567	23	45	36	m	
8	32156	67	72	81	m	
9	33456	81	77	88	f	
10	67945	74	91	92	f	
11						

Display I.4: The `marks` worksheet.

Data arises from the process of taking measurements of variables in some real-world context. For example, in a population of students, suppose that we are conducting a study of academic performance in a Statistics course. Specifically, suppose that we want to examine the relationship between grades in Statistics, grades in a Calculus course, grades in a Physics course, and gender. So we collect the following information for each student in the study: student number, grade in Statistics, grade in Calculus, grade in Physics, and gender. Therefore, we have five variables—student number and the grades in the three subjects are *numeric variables*, and gender is a *text variable*. Let us further suppose that there are ten students in the study.

Display I.4 gives a possible outcome from collecting the data in such a study. Column C1 contains the student number (note that this is a categorical variable even though it is a number). The student number primarily serves as an identifier so that we can check that the data has been entered correctly. This is something you should always do as a first step in your analysis. Columns C2–C4 contain the student grades in their Statistics, Calculus, and Physics courses and column C5 contains the gender data. Notice that a column contains the values collected for a single variable, and a row contains the values of all the variables for a single student. Sometimes, a row is referred to as an *observation* or *case*. Observe that the data for this study occupies a 10×5 subtable of the full worksheet. All of the other blank entries of the worksheet can be ignored, as they are undefined.

There will be limitations on the number of columns and rows you can have in your worksheet, and this depends on the particular implementation of Minitab

you are using. So if you plan to use Minitab for a large problem, you should check with the system person or further documentation to see what these limitations are. For example, in Minitab 15 there is a limitation of 4000 columns and 10^7 rows.

Associated with a worksheet is a table of *constants*. Typically, these are numbers that you want to use in some arithmetical operation applied to every value in a column. For example, you may have recorded heights of people in inches and want to convert these to heights in centimeters. So you must multiply every height by the value 2.54. The Minitab constants are labeled K1, K2, etc. To continue with the above problem, we might assign the value 2.54 to K1. In Section I.7.4, we show how to make such an assignment, and in Section I.10.1 we show how to multiply every entry in a column by this value.

There is an additional structure in Minitab beyond the worksheet called the *project*. A project can have multiple worksheets associated with it. Also, a project can have associated with it various graphs and records of the commands you have typed and the output obtained while working on the worksheets. Projects, which are discussed in Appendix A, can be saved and retrieved for later work.

6 Minitab Commands

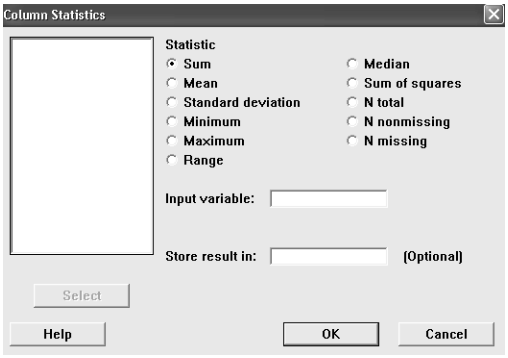
We will now begin to introduce various Minitab commands to get data into a worksheet, edit a worksheet, perform various operations on the elements of a worksheet, and save and access a saved worksheet. Before we do, however, it is useful to know something about the basic structure of all Minitab commands. Associated with every command is, of course, its *name,* as in File ▶ Exit and Help. Most commands also take *arguments,* and these arguments are column names, constants, and sometimes file names.

Commands can be accessed by making use of the File, Edit, Data, Calc, Stat, Graph, and Editor entries in the menu bar. Clicking any of these brings up a list of commands that you can use to operate on your worksheet. The lists that appear may depend on which window is active; for example, either a Data window or the Session window. Unless otherwise specified, we will always assume that the Session window is active when discussing menu commands. If a command name in a list is faded, then it is not available.

Typically, using a command from the menu bar requires the use of a *dialog box* or *dialog window* that opens when you click on a command in the list. These are used to provide the arguments and subcommands to the command and specify where the output is to go. Dialog boxes have various boxes that must be filled in to correctly execute a command. Clicking in a box that needs to be filled in typically causes a *variable list* of all items in the active worksheet that can be placed in that box to appear in the left-most box. Double clicking on items in the variable list places them in the box, or, alternatively, you can type them in directly. When you have filled in the dialog box and clicked OK, the command is printed in the Session window and executed. Any output is

also printed in the Session window. Dialog boxes have a Help button that can be used to learn how to make the entries.

For example, suppose that we want to calculate the *mean* of column C2 in the worksheet `marks`. Then the command Calc ▶ Column Statistics brings up the dialog box shown in Display I.5. Notice that the radio button Sum is filled in. Clicking the radio button labeled Mean results in this button being filled in and the Sum button becoming empty. Whichever button is filled in will result in that statistic being calculated for the relevant columns when we finally implement the command by clicking OK.

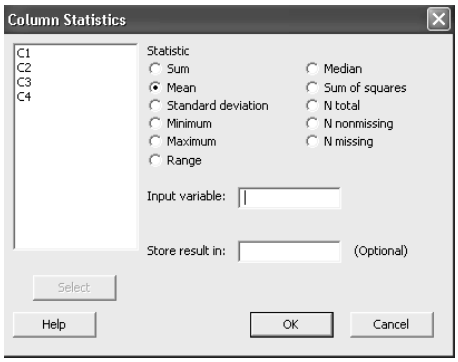


Display I.5: Initial view of the dialog box for Column Statistics.

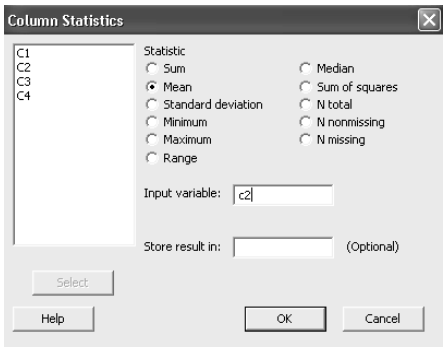
Currently, there are no columns selected, but clicking in the Input variable box brings up a list of possible columns in the display window on the left. The results of these operations are shown in Display I.6. We double click on C2 in the variable list, which places this entry in the Input variable box as shown in Display I.7. Alternatively, we could have simply typed this entry into the box. After clicking the OK button, we obtain the output

```
Mean of C2 = 69.9
```

in the Session window.



Display I.6: View of the dialog box for Column Statistics after selecting Mean and bringing up the variable list.



Display I.7: Final view of the dialog box for Column Statistics.

Quite often, it is faster and more convenient to simply type your commands directly into the Session window. Sometimes, it is necessary to use the Session window approach. So we now describe the use of commands in the Session window.

The basic structure of such a command with n arguments is

command name E_1,E_2,...,E_n

where E_i is the ith argument. Alternatively, we can type

command name E_1 E_2 ... E_n

if we don't want to type commas. Conveniently, if the arguments E_1,E_2,...,E_n are consecutive columns in the worksheet, we have the following short-form

command name E_1-E_n

which saves even more typing and accordingly decreases our chance of making a typing mistake. If you are going to type a long list of arguments and you don't want them all on the same line, then you can type the *continuation symbol* & where you want to break the line and then hit Enter. Minitab responds with the prompt

CONT>

and you continue to type argument names. The command is executed when you hit Enter after an argument name without a continuation character following it.

Many commands can, in addition, be supplied with various subcommands that alter the behavior of the command. The structure for commands with subcommands is

command name E_1 ... E_{n_1};
subcommand name E_{n_1+1} ... E_{n_2};
$\vdots$
subcommand name $\mathrm{E}_{n_{k-1}+1}$ ... E_{n_k}.

Notice that when there are subcommands each line ends with a semicolon until the last subcommand, which ends with a period. Also, subcommands may have

arguments. When Minitab encounters a line ending in a semicolon it expects a subcommand on the next line and changes the prompt to

```
SUBC >
```

until it encounters a period, whereupon it executes the command. If while typing in one of your subcommands you suddenly decide that you would rather not execute the subcommand—perhaps you realize something was wrong on a previous line—then type **abort** after the `SUBC >` prompt and hit Enter. As a further convenience, it is worth noting that you need to only type in the first four letters of any Minitab command or subcommand.

For example, to calculate the mean of column C2 in the worksheet `marks`, we can use the **mean** command in the Session window, as in

```
MTB > mean c2
```

and we obtain the same output in the Session window as before.

There are additional ways in which you can input commands to Minitab. Instead of typing the commands directly into the Session window, you can also type these directly into the Command Line Editor, which is available via Edit ▶ Command Line Editor. Multiple commands can then be typed directly into a box that pops up and executed when the Submit Commands button is clicked. Output appears in the Session window. Also, many commands are available on a *toolbar* that lies just below the menu bar at the top of the Minitab window. There is a different toolbar depending upon which window is active. We give a brief discussion of some of the features available in the toolbar in later sections.

7 Entering Data into a Worksheet

There are various methods for entering data into a worksheet. The simplest approach is to use the *Data window* to enter data directly into the worksheet by clicking your mouse in a cell and then typing the corresponding data entry and hitting Enter. Remember that you can make a Data window active by clicking anywhere in the window or by using Window in the menu bar. If you type any character that is not a number, Minitab automatically identifies the column containing that cell as a text variable and indicates that by appending T to the column name, for example, C5-T in Display I.4. You do not need to append the T when referring to the column. Also, there is a *data direction arrow* in the upper-left corner of the data window that indicates the direction the cursor moves after you hit Enter. Clicking on it alternates between row-wise and column-wise data entry. Certainly, this is an easy way to enter data when it is suitable. Remember, columns are variables and rows are observations! Also, you can have multiple data windows open and move data between them. Use the command File ▶ New to open a new worksheet.

7.1 Importing Data

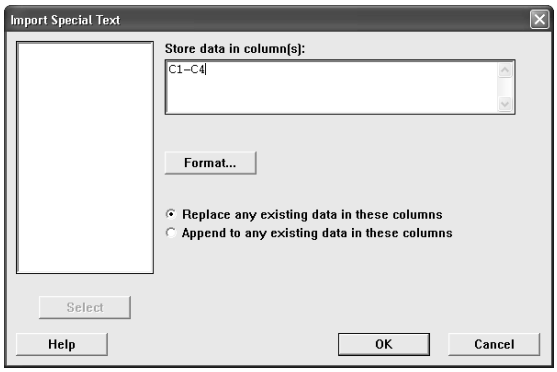
If your data is in an external file (not an `.mtw` file), you will need to use File ▶ Other Files ▶ Import Special Text to get the data into your worksheet. For example, suppose in the file `marks.txt` we have the following data recorded, just as it appears.

```
12389 81 85 78
97658 75 72 62
53546 77 83 81
55542 63 42 55
11223 71 82 67
77788 87 56 *
44567 23 45 35
32156 67 72 81
33456 81 77 88
67945 74 91 92
```

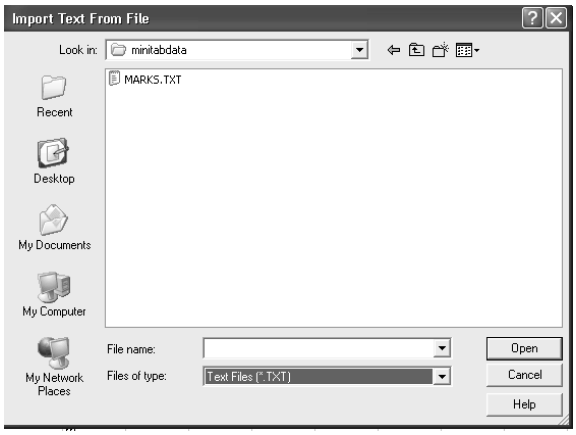
Each row corresponds to an observation, with the student number being the first entry, followed by the marks in the student's Statistics, Calculus, and Physics courses. These entries are separated by blanks.

Notice the * in the sixth row of this data file. In Minitab, a * signifies a *missing numeric value,* i.e., a data value that for some reason is not available. Alternatively, we could have just left this entry blank. A *missing text value* is simply denoted by a blank. Special attention should be paid to missing values. In general, Minitab statistical analyses ignore any cases (observations) that contain missing data except that the output of the command will tell you how many cases were ignored because of missing data. It is important to pay attention to this information. If your data is riddled with a large number of missing values, your analysis may be based on very few observations—even if you have a large data set!

When data in such a file is *blank-delimited* like this, it is very easy to read in. After the command File ▶ Other Files ▶ Import Special Text, we see the dialog box shown in Display I.8 less C1–C4 in the Store data in column(s): box. We typed `C1-C4` into this window to indicate that we want the data read in to be stored in these columns. Note that it doesn't matter if we use lower-case or upper-case for the column names, as Minitab is not case sensitive. After clicking OK, and navigating to the Windows folder `C:\minitabdata`, we see the dialog box depicted in Display I.9, which we use to indicate from which file we want to read the data. Note that if your data is in `.txt` files rather than `.dat` files, you will have to indicate that you want to see these in the Files of type box by selecting Text Files (and then all files with this suffix in the Data directory are listed) or perhaps All Files. Clicking on `marks.txt` results in the data being read into the worksheet.



Display I.8: Dialog box for importing data from external file.



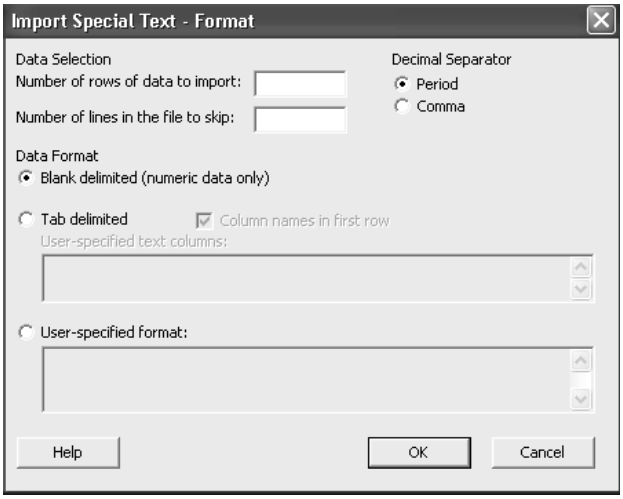
Display I.9: Dialog box for selecting file from which data is to be read in.

Of course, this data set does not contain the text variable denoting the student's gender. Suppose that the file `marksgend.txt` contains the following data exactly as typed.

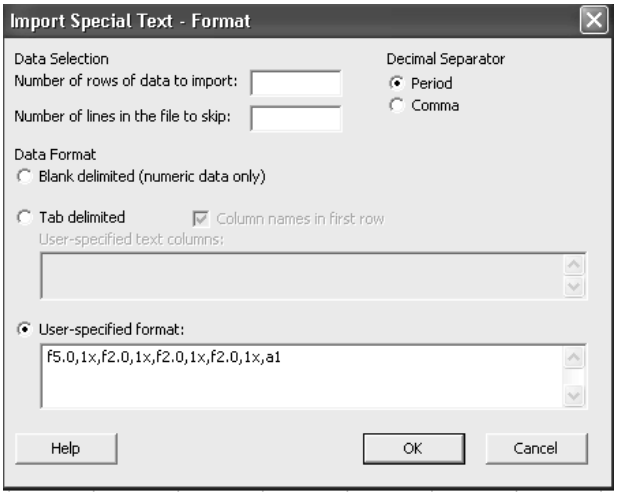
```
12389 81 85 78 m
97658 75 72 62 m
53546 77 83 81 f
55542 63 42 55 m
11223 71 82 67 f
77788 87 56  * f
44567 23 45 35 m
32156 67 72 81 m
33456 81 77 88 f
67945 74 91 92 f
```

As this file contains text data in the fifth column, we must tell Minitab how the data is *formatted* in the file. To access this feature, we click on the Format button in the dialog box shown in Display I.8. This brings up the dialog box shown in Display I.10. To indicate that we will specify the format, we click the

radio button User-specified format and fill the particular format into the box as shown in Display I.11. The format statement says that we are going to read in the data according to the following rule: a numeric variable occupying five spaces and with no decimals, followed by a space, a numeric variable occupying two spaces with no decimals, a space, a numeric variable occupying two spaces with no decimals, a space, a numeric variable occupying two spaces with no decimals, a space, and a text variable occupying one space. This rule must be rigorously adhered to or errors will occur.



Display I.10: Initial dialog box for formatted input.



Display I.11: Dialog box for formatted input with the format filled in.

So the rules you need to remember, if you use formatted input, are that `ak` indicates a text variable occupying `k` spaces, `kx` indicates `k` spaces, and `fk.l` indicates a numeric variable occupying `k` spaces, of which `l` are to the right

of the decimal point. Note if a data value does not fill up the full number of spaces allotted to it in the format statement, it must be right justified in its field. Also, if a decimal point is included in the number, this occupies one of the spaces allocated to the variable and similarly for a minus or plus sign. There are many other features to formatted input that we will not discuss here. Use the Help button in the dialog box for information on these features. Finally, clicking on the OK button reads this data into a worksheet as depicted in Display I.4. Typically, we try to avoid the use of formatted input because it is somewhat cumbersome, but sometimes we must use it.

In the session environment, the **read** command is available for inputting data into a worksheet with capabilities similar to what we have described. For example, the commands

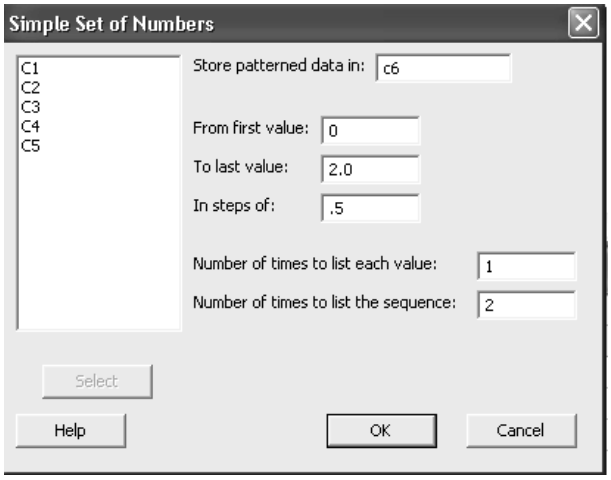
```
MTB >read c1-c4
DATA>12389 81 85 78
DATA>97658 75 72 62
DATA>53546 77 83 81
DATA>55542 63 42 55
DATA>11223 71 82 67
DATA>77788 87 56 *
DATA>44567 23 45 35
DATA>32156 67 72 81
DATA>33456 81 77 88
DATA>67945 74 91 92
DATA>end
 10 rows read.
```

place the first four columns into the `marks` worksheet. After typing `read c1-c4` after the `MTB >` prompt and hitting Enter, Minitab responds with the `DATA>` prompt, and we type each row of the worksheet in as shown. To indicate that there is no more data, we type `end` and hit Enter. Similarly, we can enter text data in this way but can't combine the two unless we use a **format** subcommand. We refer the reader to **help** for more description of how this command works.

7.2 Patterned Data

Often, we want to input *patterned data* into a worksheet. By this we mean that the values of a variable follow some determined rule. We use the command Calc ▶ Make Patterned Data for this. For example, implementing this command with the entries in the dialog box depicted in Display I.12 (for a Simple Set of Numbers) adds a column C6 to the `marks` worksheet with the sequence $0, 0.5, 1.0, 1.5, 2.0$ repeated twice. For this we entered 0 in the From first value box, a 2.0 in the To last value box, a .5 in the In steps of box, a 1 in the Number of times to list each value box, and a 2.0 in the Number of times to list the whole sequence box. Basically, we can start a sequence at any number m and successively increment this with any number $d > 0$ until the next addi-

tion would exceed the last value n prescribed, repeat each element l times, and finally repeat the whole sequence k times.



Display I.12: Dialog box for making patterned data with some entries filled in.

There is some shorthand associated with patterned data that can be very convenient. For example, typing $m : n$ in a Minitab command is equivalent to typing the values $m, m+1, \ldots, n$ when $m < n$ and $m, m-1, ..., n$ when $m > n$, and m when $m = n$. The expression $m : n/d$, where $d > 0$, expands to a list as above but with the increment of d or $-d$, whichever is relevant, replacing 1 or -1. If $m < n$, then d is added to m until the next addition would exceed n, and if $m > n$, then d is subtracted from m until the next subtraction would be lower than n. The expression $k(m : n/d)$ repeats $m : n/d$ for k times, while $(m : n/d)l$ repeats each element in $m : n/d$ for l times. The expression $k(m : n/d)l$ repeats $(m : n/d)l$ for k times.

The **set** command is available in the Session window to input patterned data. For example, suppose we want C6 to contain the ten entries 1, 2, 3, 4, 5, 5, 4, 3, 2, 1. The command

```
MTB >set c6
DATA>1:5
DATA>5:1
DATA>end
```

does this. Also, we can add elements in parentheses. For example, the command

```
MTB >set c6
DATA>(1:2/.5 4:3/.2)
DATA>end
```

creates the column with entries 1.0, 1.5, 2.0, 4.0, 3.8, 3.6, 3.4, 3.2, 3.0. The multiplicative factors k and l can also be used in such a context. Obviously, there is a great deal of scope for entering patterned data with **set**. The general syntax of the set command is

set E_1

where E_1 is a column.

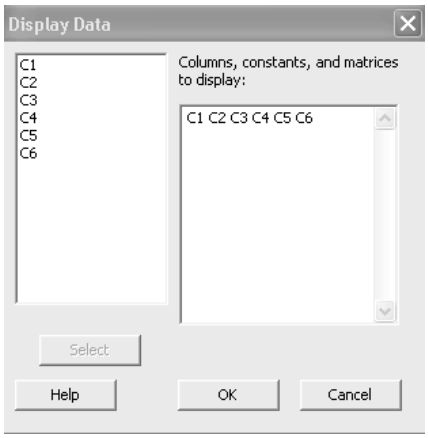
7.3 Printing Data in the Session Window

Once we have entered the data into the worksheet, we should always check that we have made the entries correctly. Typically, this means printing out the worksheet and checking the entries. The command Data ▶ Display Data will print the data you ask for in the Session window. For example, with the worksheet `marks` the dialog box pictured in Display I.13 causes the contents of this worksheet to be printed when we click on OK. We selected which variables to print by first clicking in the Columns, constants, and matrices to display box, and then double clicking on the variables in the variable list on the left.

The **print** command is available in the Session window and is often convenient to use. The general syntax for the **print** command is

print E_1 ... E_m

where E_1, ..., E_m are columns and constants. This prints the contents of these columns and constants in the Session window.

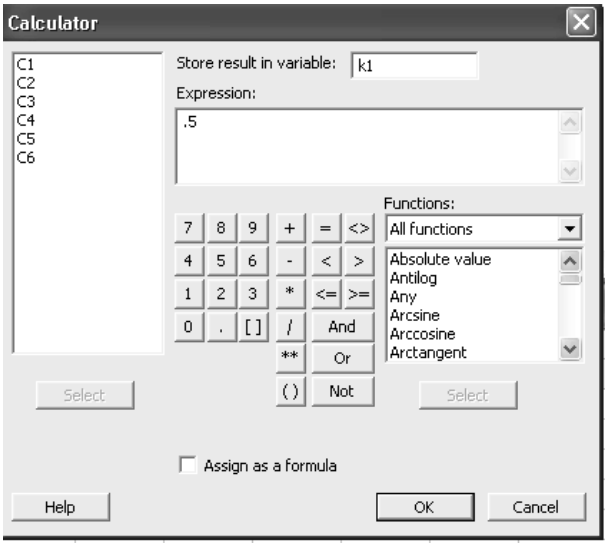


Display I.13: Dialog box for printing worksheet in the Session window.

7.4 Assigning Constants

To enter constants, we use the Calc ▶ Calculator command and fill in the dialog box appropriately. For example, suppose we want to assign the values `k1=.5`, `k2=.25`, and `k3=.25` to the constants k1, k2, and k3. These could serve as weights to calculate a weighted average of the marks in the `marks` worksheet. Then the Calc ▶ Calculator command leads to the dialog box displayed in Display I.14, where we have typed k1 into the Store result in variable box and the value .5 into the Expression box. Clicking on OK then makes the assignment.

Note that we can assign text values to constants by enclosing the text in double quotes. We will talk about further features of Calculator later in this manual. Similarly, we assign values to k2 and k3.



Display I.14: Filled in dialog box for assigning the constant k1 the value .5.

The **let** command is available in the Session window and is quite convenient. The following commands make this assignment and then we check, using the **print** command, that we have entered the constants correctly.

```
MTB >let k1=.5
MTB >let k2=.25
MTB >let k3=.25
MTB >print k1-k3
K1 0.500000
K2 0.250000
K3 0.250000
```

Also, we can assign constants text values. For example,

```
MTB >let k4="result"
```

assigns K4 the value `result`. Note the use of double quotes.

7.5 Naming Variables and Constants

It often makes sense to give the columns and constants names rather than just referring to them as C1, C2, ..., K1, K2, etc. This is especially true when there are many variables and constants, as it would be easy to slip and use the wrong column in an analysis and then wind up making a mistake. To assign a name to a variable, simply go to the blank cell at the top of the column in the worksheet corresponding to the variable and type in an appropriate name. For example,

we have used `studid, statistics, calculus, physics,` and `gender` for the names of C1, C2, C3, C4, and C5, respectively, and these names appear in Display I.15.

↓	C1	C2	C3	C4	C5-T	C6	C7
	studid	statistics	calculus	physics	gender		
1	12399	81	85	78	m		
2	97658	75	72	62	m		
3	53546	77	83	81	f		
4	55542	63	42	56	m		
5	11223	71	82	67	f		
6	77788	87	56	*	f		
7	44567	23	45	36	m		
8	32156	67	72	81	m		
9	33456	81	77	88	f		
10	67945	74	91	92	f		
11							
12							

Display I.15: Worksheet `marks` with named variables.

In the Session window, the **name** command is available for naming variables and constants. For example, the commands

```
MTB >name c1 'studid' c2 'stats' c3 'calculus' &
CONT>c4 'physics' c5 'gender' &
CONT>k1 'weight1' k2 'weight2' k3 'weight3'
```

give the names `studid` to C1, `stats` to C2, `calculus` to C3, `physics` to C4, `gender` to C5, `weight1` to K1, `weight2` to K2, and `weight3` to K3. Notice that we have made use of the continuation character & for convenience in typing in the full input to **name**. When using the variables as arguments, just enclose the names in single quotes. For example,

```
MTB >print 'studid' 'calculus'
```

prints out the contents of these variables in the Session window.

Variable and constant names can be at most 31 characters in length, cannot include the characters #, ', and cannot start with a leading blank or *. Recall that Minitab is not case sensitive, so it does not matter if we use lower-case or upper-case letters when specifying the names.

7.6 Information about a Worksheet

We can get information on the data we have entered into the worksheet by using the **info** command in the Session window. For example, we get the following results based on what we have entered into the `marks` worksheet so far.

```
MTB >info
```

```
    Column    Name        Count Missing
A   C1        studid         10       0
    C2        stats          10       0
    C3        calculus       10       0
    C4        physics        10       1
A   C5        gender         10       0
    Constant  Name       Value
    K1        weight1    0.500000
    K2        weight2    0.250000
    K3        weight3    0.250000
```

Notice that the **info** command tells us how many missing values there are and in what columns they occur and also the values of the constants.

This information can also be accessed directly from the *Project Manager window* via Window ▶ Project Manager.

7.7 Editing a Worksheet

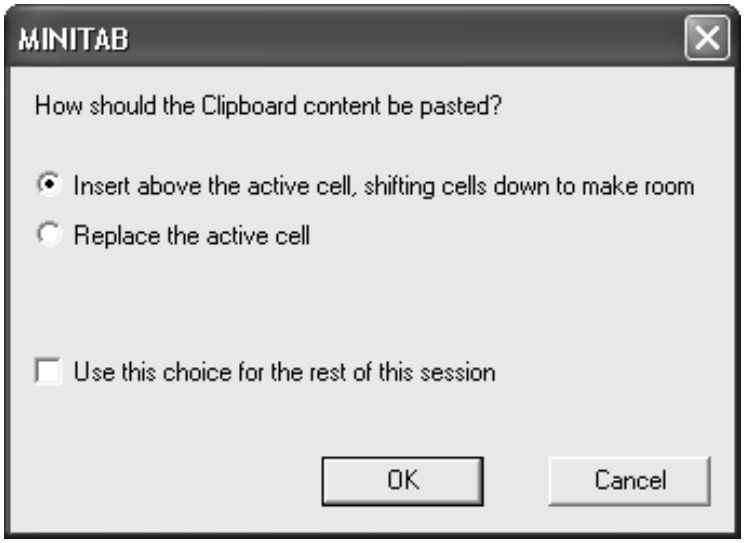
It often happens that after data entry we notice that we have made some mistakes or we obtain some additional information, such as more observations. So far, the only way we could change any entries in the worksheet or add some rows is to reenter the whole worksheet!

Editing the worksheet is straightforward because we simply change any cells by retyping their entries and hitting the Enter key. We can add rows and columns at the end of the worksheet by simply typing new data entries in the relevant cells. To insert a row before a particular row, simply click on any entry in that row and then the menu command Editor ▶ Insert Rows. Fill in the blank entries in the new row. To insert a column before a particular column, simply click on any entry in that column and then the menu command Editor ▶ Insert Columns. Fill in the blank entries in the new column. To insert a cell before a particular cell, simply click on any entry in that cell and the menu command Editor ▶ Insert Cells. Fill in the blank entry in the new cell that appears in place of the original with all other cells in that column—and only that column—pushed down.

If you wish to clear a number of cells in a block, click in the cell at the start of the block, and holding the mouse key down, drag the cursor through the block so that it is highlighted in black. Click on the Cut icon on the Minitab *taskbar,* and all the entries will be deleted. Cells immediately below the block move up to fill in the vacated places. A convenient method for clearing all the data entries in a worksheet, with the relevant Data window active, is to use the command Edit ▶ Select All Cells, which causes all the cells to be highlighted, and click on the Cut icon. Always save the contents of the current worksheet before doing this unless you are absolutely sure you don't need the data again. We discuss how to save the contents of a worksheet in Section I.8.

To copy a block of cells, click in the cell at the start of the block and, holding the mouse key down, drag the cursor through the block so that it is highlighted

in black, but, instead of hitting the backspace key, use the command Edit ▶ Copy Cells or click on the Copy icon on the Minitab taskbar. The block of cells is now copied to your clipboard. If you not only want to copy a block of cells to your clipboard but remove them from the worksheet, use the command Edit ▶ Cut Cells or the Cut icon on the Minitab taskbar instead. Note that any cells below the removed block will move up to replace these entries. To paste the block of cells into the worksheet, click on the cell before which you want the block to appear or that is at the start of the block of cells you wish to replace and issue the command Edit ▶ Paste Cells, or use the Paste icon on the Minitab taskbar. A dialog box appears as in Display I.16, where you are prompted as to what you want to do with the copied block of cells. If you feel that a cutting or pasting was in error, you can undo this operation by using Edit ▶ Undo Cut or Edit ▶ Undo Paste, respectively, or use the Undo icon on the Minitab taskbar.



Display I.16: Dialog box that determines how a block of copied cells is used.

An alternative approach is available for copying operations using Data ▶ Copy and filling in the dialog box appropriately. We refer the reader to the online manual for more description of these features.

One can also delete selected rows from specified columns using Data ▶ Delete Rows and filling in the dialog box appropriately. Notice, however, that whenever we delete a cell, the contents of the cells beneath the deleted one in that column simply move up to fill the cell. The cell entry does not become missing; rather, cells at the bottom of the column become undefined! If you delete an entire row, this is not a problem because the rows below just shift up. For example, if we delete the third row, then in the new worksheet, after the deletion, the third row is now occupied by what was formerly the fourth row. Therefore, you should be careful, when you are not deleting whole rows, to ensure that you get the result you intended.

Note that if you should delete all the entries from a column, this variable is still in the worksheet, but it is empty now. If you wish to delete a variable and all its entries, this can be accomplished from Data ▶ Erase Variables and filling in the dialog box appropriately. This is a good idea if you have a lot of variables and no longer need some of them.

There are various commands in the Session window available for carrying out these editing operations. For example, the **restart** command in the Session window can be used to remove all entries from a worksheet. The **let** command allows you to replace individual entries. For example,

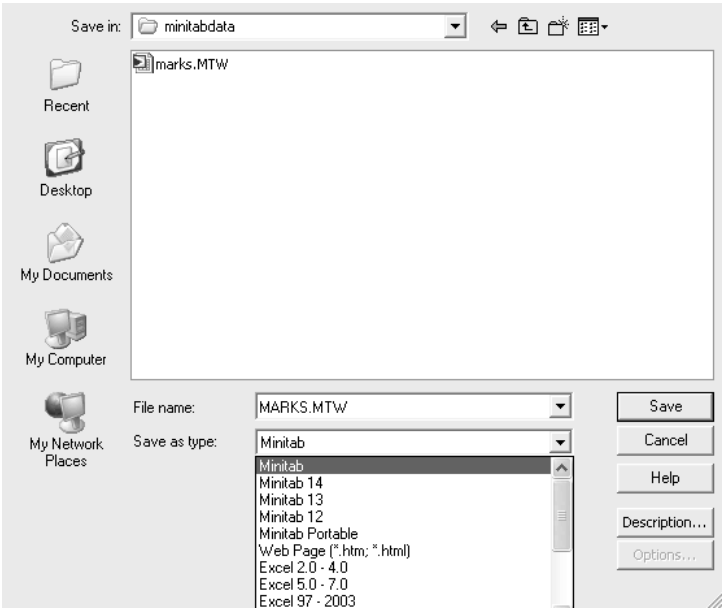
```
MTB > let c2(2)=3
```

assigns the value 3 to the second entry in the column C2. The **copy** command can be used to copy a block of cells from one place to another. The **insert** command allows you to insert rows or observations anywhere in the worksheet. The **delete** command allows you to delete rows. The **erase** command is available for the deletion of columns or variables from the worksheet. As it is more convenient to edit a worksheet by directly working on the worksheet and using the menu commands, we do not discuss these commands further here.

8 Saving, Retrieving, and Printing

Quite often, you will want to save the results of all your work in creating a worksheet. If you exit Minitab before you save your work, you will have to reenter everything. So we recommend that you always save. To use the commands of this section, make sure that the Worksheet window of the worksheet in question is active.

Use File ▶ Save Current Worksheet to save the worksheet with its current name, or the default name if it doesn't have one. If you want to provide a name or store the worksheet in a new location, then use File ▶ Save Current Worksheet As and fill in the dialog box depicted in Display I.17 appropriately. The Save in box at the top contains the name of the folder in which the worksheet will be saved once you click on the Save button. Here the folder is called `minitabdata`, and you can navigate to a new folder using the Up One Level button immediately to the right of this box. The next button allows you to create a subfolder within the current folder. The box immediately below contains a list of all files of type `.mtw` in the current folder. You can select the type of file to display by clicking on the arrow in the Save as type box, which we have done here, and click on the type of file you want to display that appears in the drop-down list. There are several possibilities including saving the worksheet in other formats, such as Excel. Currently, there is one `.mtw` file in the folder `minitabdata` and it is called `marks.mtw`. If you want to save the worksheet with a particular name, type this name in the File name box and click on the Save button.

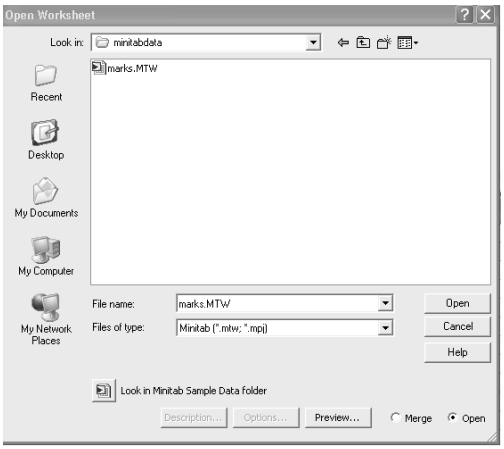


Display I.17: Dialog box for saving a worksheet.

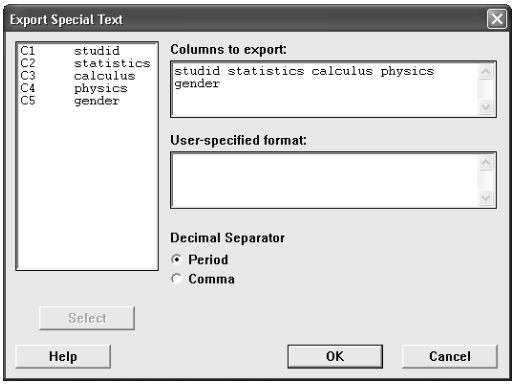
To retrieve a worksheet, use File ▶ Open Worksheet and fill in the dialog box as depicted in Display I.18 appropriately. The various windows and buttons in this dialog box work as described for the File ▶ Save Current Worksheet As command, with the exception that we now type the name of the file we want to open in the File name box, alternatively click on the relevant file, and then click on the Open button.

To print a worksheet, use the command File ▶ Print Worksheet. The dialog box that subsequently pops up allows you to control the output in a number of ways.

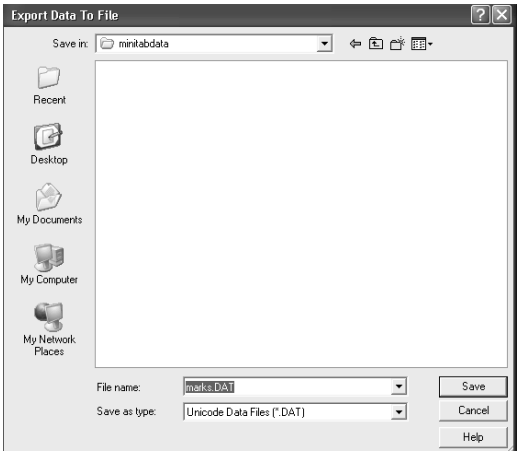
It may be that you would prefer to write out the contents of a worksheet to an external file that can be edited by an editor or perhaps used by some other program. This will not be the case if we save the worksheet as an `.mtw` file as only Minitab can read these. To do this, use the command File ▶ Other Files ▶ Export Special Text, filling in the dialog box and specifying the destination file when prompted. For example, if we want to save the contents of the `marks` worksheet, this command results in the dialog box of Display I.19 appearing. We have entered all five columns into the Columns to export box and have not specified a format, so the columns will be stored in the file with single blanks separating the columns. Clicking the OK button results in the dialog box of Display I.20 appearing. Here, we have typed in the name `marks.dat` to hold the contents. Note that while we have chosen a `.dat` type file, we also could have chosen a `.txt` type file. Clicking on the Save button results in a file `marks.dat` being created in the folder `minitabdata` with contents as in Display I.21.



Display I.18: Dialog box for retrieving a worksheet.



Display I.19: Dialog box for saving the contents of a worksheet to an external (non-Minitab) file.



Display I.20: Dialog box for selecting external file to hold contents of a worksheet.

```
12389    81    85    78  m
97658    75    72    62  m
53546    77    83    81  f
55542    63    42    55  m
11223    71    82    67  f
77788    87    56     *  f
44567    23    45    35  m
32156    67    72    81  m
33456    81    77    88  f
67945    74    91    92  f
```

Display I.21: Contents of the file `marks.dat`.

In the Session window, the commands **save** and **retrieve** are available for saving and retrieving a worksheet in the `.mtw` format and the command `write` is available for saving a worksheet in an external file. We refer the reader to **help** for a description of how these commands work.

9 Mathematical Operations

When carrying out a data analysis, a statistician is often called upon to transform the data in some way. This may involve applying some simple transformation to a variable to create a new variable—for example, take the natural logarithm of every grade in the `marks` worksheet—to combining several variables together to form a new variable—for example, calculate the average grade for each student in the `marks` worksheet. In this section, we present some of the ways of doing this.

9.1 Arithmetical Operations

Simple arithmetic can be carried out on the columns of a worksheet using the arithmetical operations of addition +, subtraction −, multiplication *, division /, and exponentiation ** via the Calc ▶ Calculator command. When columns are added together, subtracted one from the other, multiplied together, divided one by the other (make sure there are no zeros in the denominator column), or one column exponentiates another, these operations are always performed component-wise. For example, C1*C2 means that the ith entry of C1 is multiplied by the ith entry of C2, etc. Also, make sure that the columns on which you are going to perform these operations correspond to numeric variables! While these operations have the order of precedence **, */, +−, parentheses ( ) can and should be used to ensure an unambiguous result. For example, suppose in the `marks` worksheet we want to create a new variable by taking the average of the Statistics and Calculus grades and then subtracting this average from the Physics grade and placing the result in C6. Filling in the dialog box, corre-

sponding to Calc ▶ Calculator, as shown in Display I.22 accomplishes this when we click on the OK button. Note that we can either type the relevant expression into the Expression box or use the buttons and double click on the relevant columns. Further, we type the column where we wish to store the results of our calculation in the Store result in variable box. These operations are done on the corresponding entries in each column; corresponding entries in the columns are operated on according to the formula we have specified, and a new column of the same length containing all the outcomes is created. Note that the sixth entry in C6 will be * (or missing) because this entry was missing for C4.

These kinds of operations can also be carried out directly in the Session window using the **let** command, and in some ways this is a simpler approach. For example, the session command

```
MTB >let c6=c4-(c2+c3)/2
```

accomplishes this.

Display I.22: Dialog box for carrying out mathematical calculations.

We can also use these arithmetical operations on the constants K1, K2, etc., and numbers to create new constants or use the constants as *scalars* in operations with columns. For example, suppose that we want to compute the weighted average of the Statistics, Calculus, and Physics grades, where Statistics gets twice the weight of the other grades. Suppose that we created, as part of the `marks` worksheet, the constants `weight1 = .5`, `weight2 = .25`, and `weight3 = .25` in K1, K2, and K3, respectively. So this weighted average is computed via the command

```
MTB >let c7='weight1'*'stats'+'weight2'*'calculus'&
CONT>+'weight3'*'physics'
```

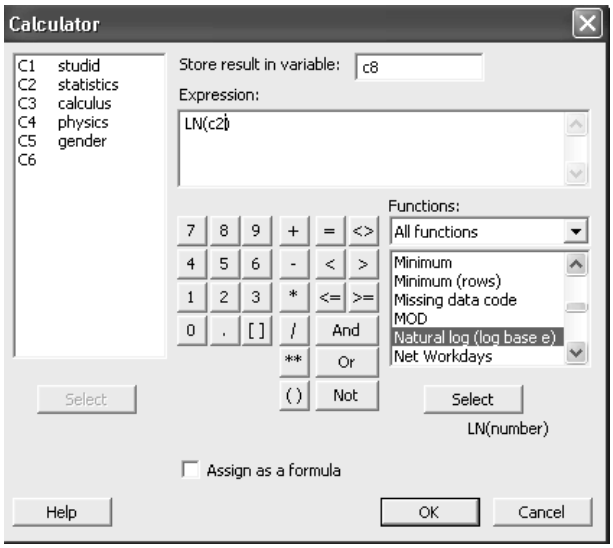
9.2 Mathematical Functions

Various mathematical functions are available in Minitab. For example, suppose we want to compute the natural logarithm of the Statistics mark for each student and store the result in C8. Using the Calc ▶ Calculator command, with the dialog box as in Display I.23, accomplishes this. A complete list of such functions is given in the Functions window when All functions is in the window directly above the list.

The same result can be obtained using the session command **let** and the natural logarithm function **ln**. For example,

```
MTB >let c8=ln(c2)
```

calculates the natural log of every entry in C2 and places the results in C8. See Appendix B.1 for a list of mathematical functions available.



Display I.23: Dialog box for mathematical calculations illustrating the use of the natural logarithm function.

9.3 Comparisons and Logical Operations

Minitab also contains the following comparison and logical operators.

Comparison operators	**Logical operators**
equal to =, **eq**	&, **and**
not equal to <>, **ne**	|, **or**
less than <, **lt**	~, **not**
greater than >, **gt**	
less than or equal to <=, **le**	
greater than or equal to >=, **ge**	

Notice that there are two choices for these operators; for example, use either the symbol >= or the mnemonic **ge.**

The comparison and logical operators are useful when we have simple questions about the worksheet that would be tedious to answer by inspection. This feature is particularly useful when we are dealing with large data sets. For example, suppose that we want to count the number of times the Statistics grade was greater than the corresponding Calculus grade in the `marks` worksheet. The command Calc ▶ Calculator gives the dialog box shown in Display I.24, where we have put `c6` in the Store result in variable box and `c2 > c3` in the Expression box. Clicking on the OK button results in the ith entry in C6 containing a 1 if the ith entry in C2 is greater than the ith entry in C3; i.e., the comparison is true, and a 0 otherwise. In this case, C6 contains the entries: 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, which the worksheet in Display I.4 verifies as appropriate. If we use Calc ▶ Calculator to calculate the sum of the entries in C6, we will have computed the number of times the Statistics grade is greater than the Calculus grade.

These operations can also be simply carried out using session commands. For example,

```
MTB >let c6=c2>c3
MTB >let k4=sum(c6)
MTB >print k4
K4 4.00000
```

accomplishes this.

The logical operators combine with the comparison operators to allow more complicated questions to be asked. For example, suppose we wanted to calculate the number of students whose Statistics mark was greater than their Calculus mark and less than or equal to their Physics mark. The commands

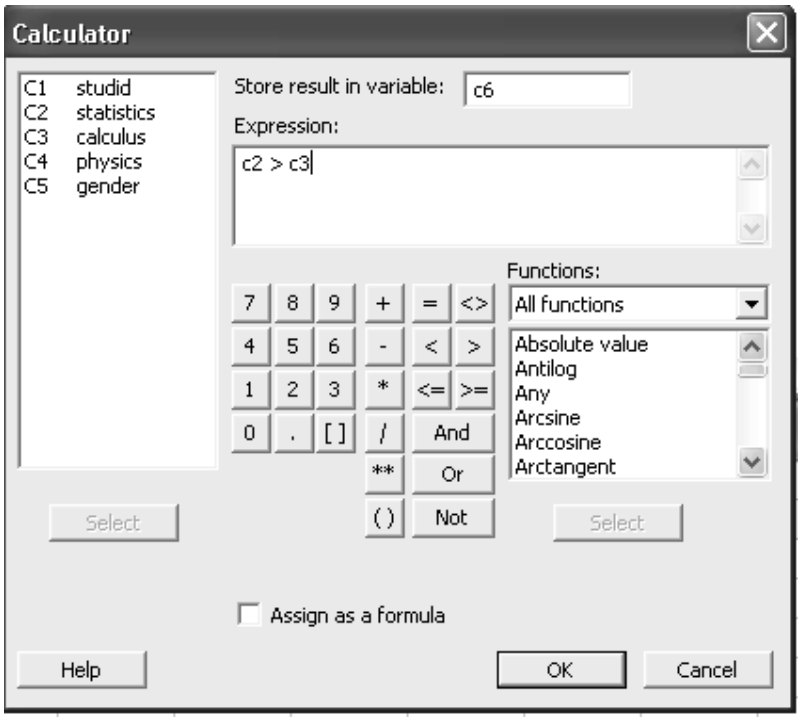
```
MTB >let c6=c2>c3 and c2<=c4
MTB >let k4=sum(c6)
MTB >print k4
K4 1.00000
```

accomplish this. In this case, both conditions `c2>c3` and `c2<=c4` have to be true for a 1 to be recorded in C6. Note that the observation with the missing Physics mark is excluded. Of course, we can also implement this using Calc ▶ Calculator and filling in the dialog box appropriately.

Text variables can be used in comparisons where the ordering is alphabetical. For example,

```
MTB >let c6=c5<"m"
```

puts a 1 in C6 whenever the corresponding entry in C5 is alphabetically smaller than `m`.

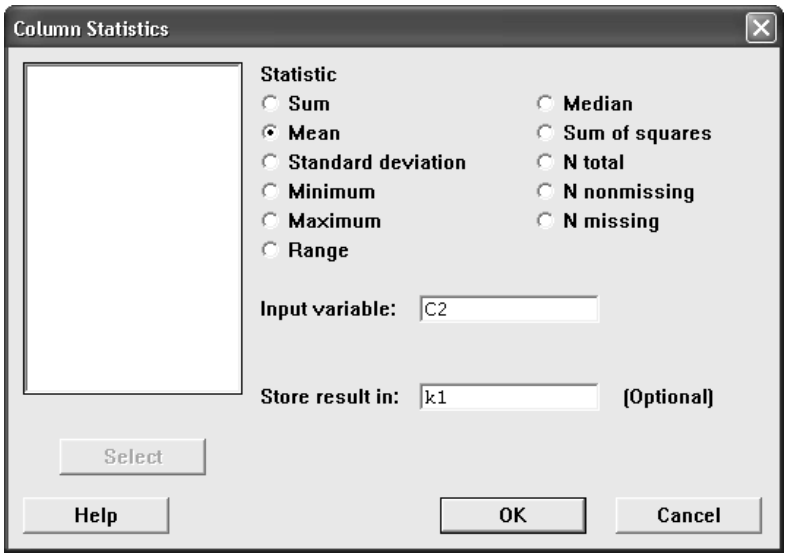


Display I.24: Dialog box for comparisons.

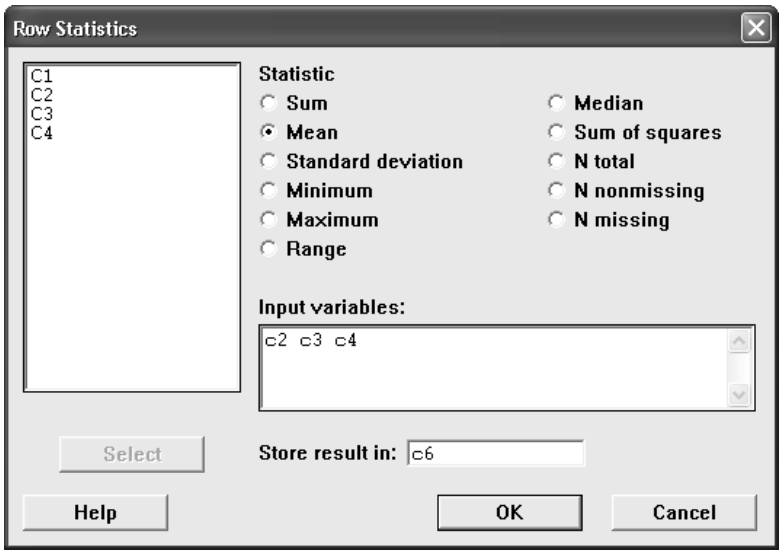
9.4 Column and Row Statistics

There are various *column statistics* that compute a single number from a column by operating on all of the elements in a column. For example, suppose that we want the mean of all the Statistics marks, i.e., the mean of all the entries in C2. The command Calc ▶ Column Statistics produces the dialog box of Display I.25, where we have selected Mean as the particular statistic to compute and C2 as the column to use. Clicking OK causes the mean of column C2 to be printed in the Session window. If we want to, we can store this result in a constant or column by making an appropriate entry in the Store result in box. In Display I.25, we see that we have stored the mean of C2 in the constant K1. We also see from the dialog box that there are a number of possible statistics that can be computed.

We can also compute statistics row-wise. One difference with column statistics is that these must be stored. For example, suppose we want to compute the average of the Statistics, Calculus, and Physics marks for each individual. The command Calc ▶ Row Statistics produces the dialog box shown in Display I.26, where we have placed C2, C3, and C4 into the Input variables box and C6 into the Store result in box.



Display I.25: Dialog box for computing column statistics.



Display I.26: Dialog box for computing row statistics.

It is also possible to compute column statistics using session commands. For example,

```
MTB >mean(c2)
MEAN = 69.900
```

computes the mean of c2. If we want to save the value for subsequent use, then the command

```
MTB >let k1=mean(c2)
```

does this. The general syntax for column statistic commands is

column statistic name(E_1)

where the operation is carried out on the entries in column E_1, and output is written to the screen unless it is assigned to a constant using the **let** command. See Appendix B.2 for a list of all the column statistics available.

Also, for most column statistics there are versions that compute *row statistics,* and these are obtained by placing **r** in front of the column statistic name. For example,

```
MTB >rmean(c2 c3 c4 c6)
```

computes the mean of the corresponding entries in C2, C3, and C4 and places the result in C6. The general syntax for row statistic commands is

row statistic name(E_1 ... E_m E_{m+1})

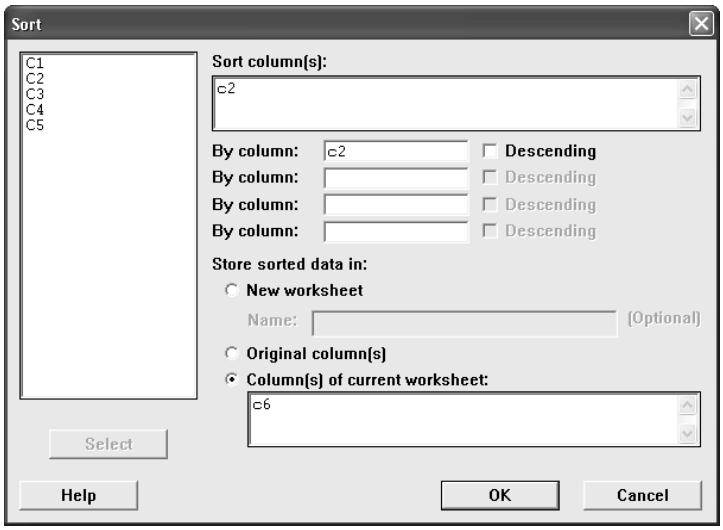
where the operations are carried out on the rows in columns E_1, ..., E_m, and the output is placed in column E_{m+1}. See Appendix B.3 for a list of all the row statistics available.

9.5 Sorting Data

It often arises that we want to *sort* a column so that its values ascend from smallest to largest or descend from largest to smallest. Note that ordering here could refer to numerical order or alphabetical order, so we also consider ordering text columns. Also, we may want to sort all the rows contained in some subset of the columns in the worksheet *by* a particular column. The Data ▶ Sort command allows us to carry out these tasks.

For example, suppose that we want to sort the entries in C2 in the `marks` worksheet—the Statistics grades—from smallest to largest and place the sorted values in C6. Then the Data ▶ Sort command brings up the dialog box shown in Display I.27, where the Sort column(s) box contains the column C2 to be sorted, the Store sorted data in box contains C6, where we will store the sorted column, and C2 is also placed in the By column box. This command results in C6 containing 23, 63, 67, 71, 74, 75, 77, 81, 81, and 87. If we had clicked the Descending box, the order of appearance of these values in C6 would have been reversed.

If we had placed another column in the By column box, say C5, then C5 would have been sorted with the values in C2 carried along and placed in C6; i.e., the values in C2 would be sorted *by* the values in C5. So all the Statistics marks of females, in the order they appear in C2, will appear in C6 first and then the Statistics marks of males. So, replacing C2 by C5 in this box would result in the values in C6 becoming 77, 71, 87, 81, 74, 81, 75, 63, 23, and 67. If we fill in the next By column box with another column, say C3, then the values in C2 are sorted first by gender and then within gender by the values in C3.



Display I.27: Dialog box for sorting.

The general syntax of the corresponding session command **sort** is

sort E_1 $E_2 \ldots E_m$ $E_{m+1} \ldots E_{2m}$

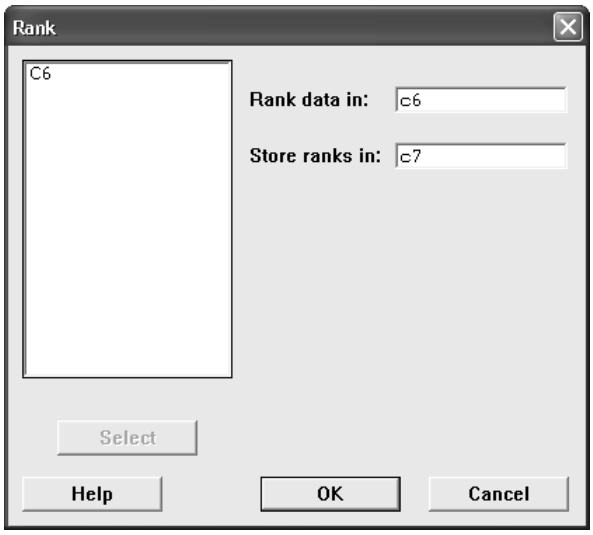
where E_1 is the column to be sorted, and E_2, ..., E_m are carried along with the results placed in columns E_{m+1}, ..., E_{2m}. Note that this sort can also be accomplished using the **by** subcommand, where the general syntax is

sort E_1 $E_2 \ldots E_m$ $E_{m+1} \ldots E_{2m}$;
by $E_{2m+1} \ldots E_n$.

where now we sort by columns E_{2m+1}, ..., E_n, sorting first by E_{2m+1}, then E_{2m+2}, etc., carrying along E_1, ..., E_m and placing the result in E_{m+1}, ..., E_{2m}. The **descending** subcommand can also be used to indicate which sorting variables we want to use in descending order, rather than ascending order.

9.6 Computing Ranks

Sometimes, we want to compute the *ranks* of the numeric values in a column. The rank r_i of the ith value in a column is a value that reflects its relative size in the column. For example, if the ith value is the smallest value, then $r_i = 1$; if it is the third smallest, then $r_i = 3$, etc. If values are the same, i.e., *tied,* then each value receives the average rank. To calculate the ranks of the entries in a column, we use the Data ▶ Rank command. For example, suppose that C6 contains the values 6, 4, 3, 2, 3, and 1. Then the Data ▶ Rank command brings up the dialog box in Display I.28, which is filled in so that the ranks of the entries in C6 are placed in C7. In this case, the ranks are 6.0, 5.0, 3.5, 2.0, 3.5, and 1.0, respectively.



Display I.28: Dialog box for computing ranks.

The syntax of the corresponding session command **rank** is

rank E_1 E_2

where E_1 is the column whose ranks we want to compute and E_2 is the column that will hold the computed ranks.

10 Exercises

1. Start Minitab and set it up so that you can type commands in the Session window and edit your output. Print the contents of the Session window.

2. Use the online manual to read and print the entry on how you can get help in Minitab.

3. Invoke the Calc ▶ Calculator command, place k1 in the Store result in variable box, read Help in the dialog box, and from this figure out how to compute the expression 203*(10345-678)/3.6. Finally, invoke the session command `print k1` and print the Session window.

4. The following data give the High and Low trading prices in dollars for various stocks on a given day on an exchange. Create a worksheet, giving the columns the same variable names. Print the worksheet to check that you have successfully entered it. Save the worksheet giving it the name `stocks`.

Stock	High	Low
ACR	7.95	7.80
MGI	4.75	4.00
BLD	112.25	109.75
CFP	9.65	9.25
MAL	8.25	8.10
CM	45.90	45.30
AZC	1.99	1.93
CMW	20.00	19.00
AMZ	2.70	2.30
GAC	52.00	50.25

5. Generate a column C1 containing all the values starting at 1 to 10 in increments of .1. Generate a column C2 containing the sequence 1:10 repeated ten times. Save these two columns in a file `columns.txt` and print this file.

6. Create a `.txt` file containing the data in Exercise 4. Using a format statement, input these data into a worksheet. Print the contents of your session.

7. Retrieve the worksheet `stocks` created in Exercise 4. Change the `Low` value in the stock MGI to 3.95. Calculate the average of the `High` and `Low` prices for all the stocks, and save this in a column called `average`. Calculate the average of all the `High` prices, and save this in a constant called `avhi`. Similarly, do this for all the `Low` prices, and save this in a constant called `avlo`. Save the worksheet using the same name. Write all the columns out to a file called `stocks.dat`. Print the file `stocks.dat` on your system printer.

8. Retrieve the worksheet created in Exercise 7. Using Minitab commands, calculate the number of stocks in the worksheet whose `average` is greater than \$5.00 and less than or equal to \$45.00.

9. Using the worksheet created in Exercise 7, insert the following stocks at the beginning of the worksheet.

Stock	High	Low
CLV	1.85	1.78
SIL	34.00	34.00
AC	14.45	14.05

   Delete the variable `average`. Print and save the worksheet.

10. (a) Using patterned data input, place the values from -10 to 10 in increments of .1 in C1.

(b) For each of the values in C1, calculate the value of the quadratic polynomial $2x^2 + 4x - 3$ (i.e., substitute the value in each entry in C1 into this expression) and place these values in C2.

(c) Using Minitab commands and the values in C1 and C2, estimate the point in the range from -10 to 10 where this polynomial takes its smallest value and what this smallest value is. (Hint: Compute the ranks of the values in C2.)

(d) Using Minitab commands and the values in C1 and C2, estimate the points in the range from -10 to 10 where this polynomial is closest to 0.

11. (a) Using patterned data input, place values in the range from 0 to 5 using an increment of .01 in C1.

    (b) Calculate the value of $1 - e^{-x}$ for each value in C1 and place the result in C2.

    (c) Using Minitab commands, find the largest value in C1 where the corresponding entry in C2 is less than or equal to .5. Note that e^{-x} corresponds to the **exp** command (see Appendix B.1) evaluated at $-x$.

12. Using patterned data input, place values in the range from -4 to 4 using an increment of .01 in C1. Calculate the value of

$$\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

    for each value in C1, and place the result in C2, where $\pi = 3.1415927$. Using **parsums** (see Appendix B.1), calculate the partial sums for C2, and place the result in C3. Multiply C3 times .01. Find the largest value in C1 such that the corresponding entry in C3 is less than or equal to .25.

Part II

Minitab for Data Analysis

Chapter 1

Looking at Data: Exploring Distributions

New Minitab commands discussed in this chapter

Calc ▶ Probability Distributions ▶ Normal
Data ▶ Code
File ▶ Open Graph
File ▶ Save Graph As
Graph ▶ Boxplot
Graph ▶ Chart
Graph ▶ Dotplot
Graph ▶ Histogram
Graph ▶ Pie Chart
Graph ▶ Probability Plot
Graph ▶ Stem-and-Leaf
Stat ▶ Basic Statistics ▶ Display Descriptive Statistics
Stat ▶ Basic Statistics ▶ Store Descriptive Statistics
Stat ▶ Tables ▶ Tally

This chapter is concerned with the various ways of presenting and summarizing a data set. By presenting data, we mean convenient and informative methods of conveying the information contained in a data set. There are two basic methods for presenting data, namely graphically and through tabulations. Still, it can be hard to summarize exactly what these presentations are saying about the data. So the chapter also introduces various summary statistics that are commonly used to convey meaningful information in a concise way.

All of these topics can involve much tedious, error-prone calculation, if we were to insist on doing them by hand. An important point is that you should almost never rely on hand calculation in carrying out a data analysis. Not only

are there many far more important things for you to be thinking about, as the text discusses, but you are also likely to make an error. On the other hand, never blindly trust the computer! Check your results and make sure that they make sense in light of the application. For this, a few simple hand calculations can prove valuable. In working through problems, you should try to use Minitab as much as possible, as this will increase your skill with the package and inevitably make your data analyses easier and more effective.

1.1 Tabulating and Summarizing Data

If a variable is categorical, we construct a table using the values of the variable and record the *frequency* (count) of each value in the data and perhaps the *relative frequency* (proportion) of each value in the data as well. These relative frequencies then serve as a convenient summarization of the data.

If the variable is quantitative, we typically *group* the data in some way; i.e., divide the range of the data into nonoverlapping intervals and record the frequency and proportion of values in each interval. Grouping is accomplished using the Data ▶ Code command discussed in Appendix C.1.

If the values of a variable are *ordered,* we can record the *cumulative distribution,* namely, the proportion of values less than or equal to each value. Quantitative variables are always ordered but sometimes categorical variables are as well; for example, when a categorical variable arises from grouping a quantitative variable.

Often, it is convenient with quantitative variables to record the *empirical distribution function,* which for data values $x_1, \ldots, x_n$ is given by $\hat{F}(x) =$ (# of $x_i \leq x)/n$ at a value x; i.e., $\hat{F}(x)$ is the proportion of data values less than or equal to x. We can summarize such a presentation via the calculation of a few quantities, such as the *first quartile,* the *median,* and the *third quartile,* or present the *mean* and the *standard deviation.*

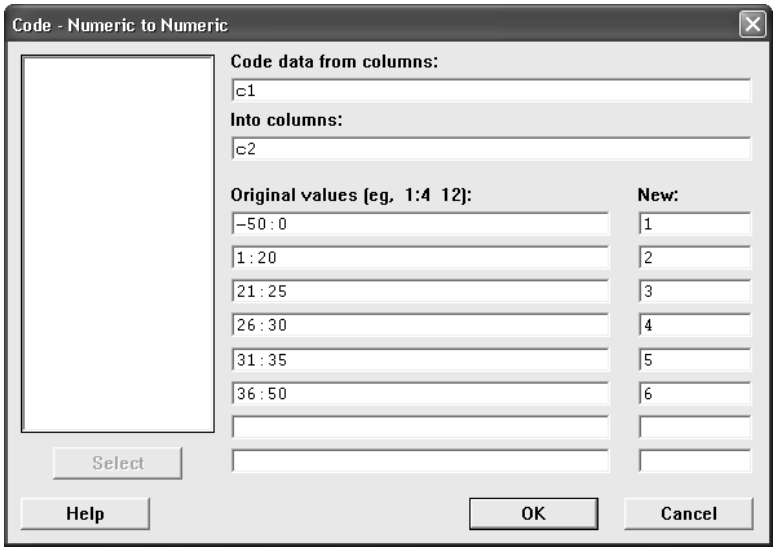
We introduce some new commands to carry out the necessary computations using the data shown in Table 1.1.1. This is data collected by A.A. Michelson and Simon Newcomb in 1882 concerning the speed of light. We will refer to these hereafter as Newcomb's data and place them in the column C1 with the name `time` in the worksheet called `newcomb`.

28	26	33	24	34	−44	27	16	40	−2	29
22	24	21	25	30	23	29	31	19	24	20
36	32	36	28	25	21	28	29	37	25	28
26	30	32	36	26	30	22	36	23	27	27
28	27	31	27	26	33	26	32	32	24	39
28	24	25	32	25	29	27	28	29	16	23

Table 1.1.1: Newcomb's data.

1.1.1 Tallying Data

The Stat ▶ Tables ▶ Tally command tabulates data. Consider Newcomb's measurements in Table 1.1.1. These data range from -44 to 40 (use minimum and maximum in Calc ▶ Calculator to calculate these values). Suppose we decide to group these into the intervals $(-50, 0]$, $(0, 20]$, $(20, 25]$, $(25, 30]$, $(30, 35]$, $(35, 50]$. Next, we want to record the frequencies, relative frequencies, cumulative frequencies, and cumulative distribution of this grouped variable. First, we used the Data ▶ Code ▶ Numeric to Numeric command, as described in Appendix C.1, to recode the data so that every value in $(-50, 0]$ is given the value 1, every value in $(0, 20]$ is given the value 2, etc., and these values are placed in C2. The dialog box for doing this is shown in Display 1.1.1.

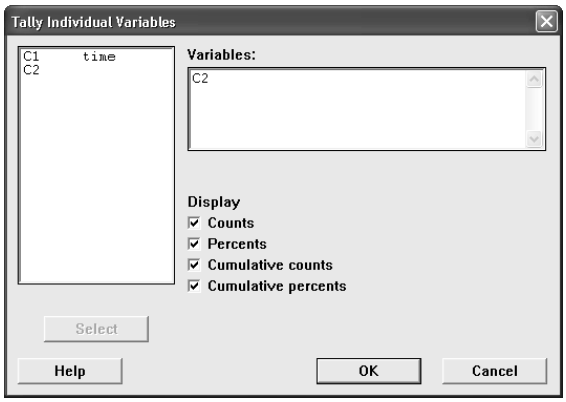


Display 1.1.1: Dialog box for recoding Newcomb's data.

Next, we used the Stat ▶ Tables ▶ Tally Individual Variables command, with the dialog box shown in Display 1.1.2, to produce the output

```
C2 Count Percent CumCnt CumPct
1      2   3.03       2   3.03
2      4   6.06       6   9.09
3     17   25.76     23  34.85
4     26   39.39     49  74.24
5     10   15.15     59  89.39
6      7   10.61     66 100.00
   N= 66
```

in the Session window.



Display 1.1.2: Dialog box for tallying the variable C2 in the `newcomb` worksheet.

We can also use the Stat ▶ Tables ▶ Tally Individual Variables command to compute the *empirical distribution function* of C1 in the `newcomb` worksheet. First, we must sort the values in C1, from smallest to largest, using the Data ▶ Sort command described in Section I.10.6, and then we apply the Stat ▶ Tables ▶ Tally Individual Variables command to this sorted variable. Note that if values are repeated, then the value of the empirical cdf at this point is the largest proportion.

The general syntax of the corresponding session command **tally** is

tally $E_1 \ldots E_m$

where E_1, ..., E_m are columns of categorical variables, and the command is applied to each column. If no subcommands are given, then only frequencies are computed, while the subcommands **percents** computes relative frequencies, **cumcnts** computes the cumulative frequency function, and **cumpcts** computes the cumulative distribution of C2. Any of the subcommands can be dropped. For example, the commands

```
MTB >sort c1 c3
MTB >tally c3;
SUBC>cumpcnts;
SUBC>store c4 c5.
```

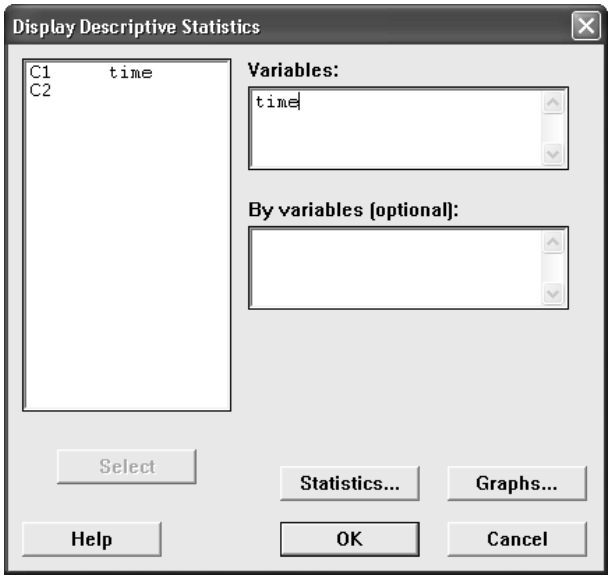
first use the **sort** command to sort the data in C1 from smallest to largest and place the results in C3. The cumulative distribution is computed for the values in C3 with the unique values in C3 stored in C4 and the cumulative distribution at each of the unique values stored in C5 via the **store** subcommand to **tally.**

1.1.2 Describing Data

The Stat ▶ Basic Statistics ▶ Display Descriptive Statistics command is used with quantitative variables to present a numerical summary of the variable values. These values are in a sense a summarization of the empirical distribution of the variable. For example, in the `newcomb` worksheet the dialog box shown in Display 1.1.3 leads to the output

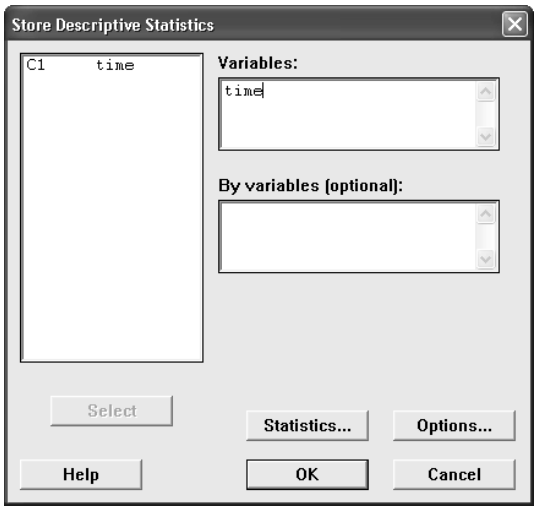
```
Variable N  N*  Mean SE Mean  StDev  Minimum   Q1   Median
time     66 0  26.21  1.32    10.75   -44.00  24.00  27.00
         Q3   Maximum
       31.00   40.00
```

in the Session window. This provides the count N, the number of missing values `N*`, the mean, standard error of the mean, standard deviation, minimum, first quartile `Q1`, median, third quartile `Q3,` and maximum of the variable C1. If we want such a summary of a variable by the values of another variable, we place these variables in the By variables box. For example, we might want such a summary for each of the groups we created in Section 1.1.1, and so we would place C2 in this box. Note that a number of summary statistics can also be computed using the Calc ▶ Column Statistics command discussed in Section I.10.3.

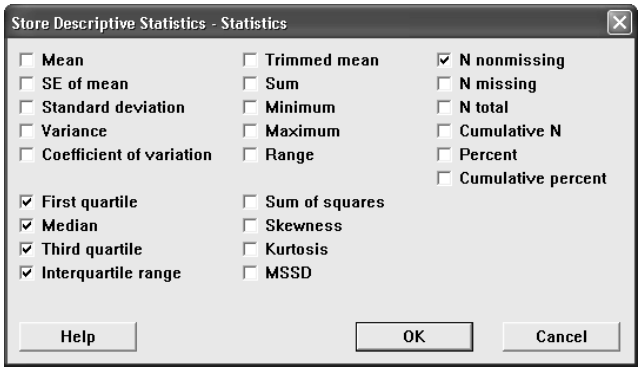


Display 1.1.3: Dialog box for computing basic descriptive statistics of a quantitative variable.

If we wish to compute some basic statistics and store these values for later use, then the Stat ▶ Basic Statistics ▶ Store Descriptive Statistics command is available for this. For example, with the `newcomb` worksheet this command leads to the dialog box shown in Display 1.1.4. Clicking on the Statistics button results in the dialog box of Display 1.1.5, where we have checked First quartile, Median, Third quartile, Inter_quartile range, and N nonmissing as the statistics we want to compute. The result of these choices is that the next available variables in the worksheet contain these values. So in this case, the values of C3–C7 are as depicted in Display 1.1.6. Note that these variables are now named as well. Note that many more statistics are available using this command.



Display 1.1.4: Dialog box for computing and storing various descriptive statistics.



Display 1.1.5: Dialog box for choosing the descriptive statistics to compute and store.

C2	C3	C4	C5	C6
Q1_1	Median1	Q3_1	IQR1	N1
24	27	31	7	66

Display 1.1.6: Values obtained for descriptive statistics using dialog boxes in Displays 1.1.4 and 1.1.5.

The general syntax of the Session command **describe,** corresponding to Stat ▶ Basic Statistics ▶ Display Descriptive Statistics, is

describe $E_1 \ldots E_m$

where E_1, ..., E_m are columns of quantitative variables and the command is applied to each column. A **by** subcommand can also be used. The **stats** command is available in the Session window if we want to store the values of statistics. We refer the reader to **help** for a description of this command.

1.2 Plotting Data

One of the most informative ways of presenting data is via a plot. There are many different types of plots within Minitab, and which one to use depends on the type of variable you have and what you are trying to learn. In this section, we describe how to use the plotting features in Minitab. There are, however, many features of plotting that we will not describe. For example, there are many graphical editing capabilities that allow you to add features, such as titles or legends. We refer the reader to Help for more details on these features.

A plot in Minitab is made in a *Graph window*. You can make multiple plots and retain each Graph window until you want to delete it simply by clicking the × symbol in the upper-right-hand corner. You make any particular Graph window active by clicking in it or by using the Window command. A plot can be saved in an external file in a variety of formats, such as Minitab graph `.mgf`, bitmap `.bmp`, JPEG `.jpg`, etc., using the File ▶ Save Graph As command. If a graph has been saved in the `.mgf` format, it can be reopened using the File ▶ Open Graph command.

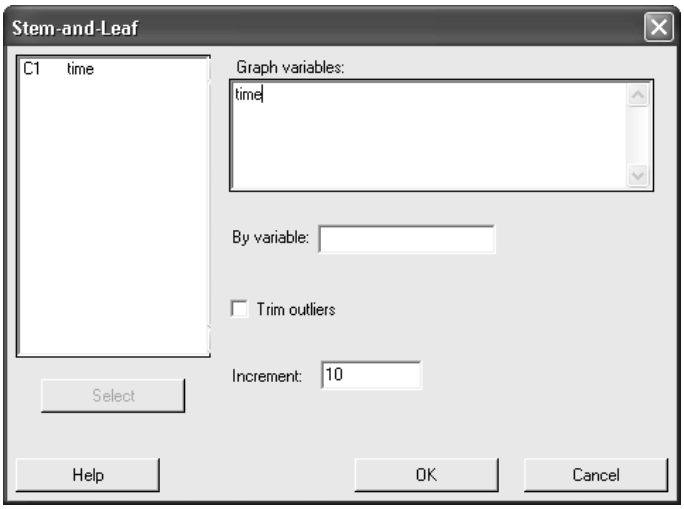
1.2.1 Stem-and-Leaf Plots

Stem-and-leaf plots are produced by the Graph ▶ Stem-and-Leaf command. These plots are also referred to as *stemplots*.

For example, using this command with the `newcomb` worksheet and the dialog box in Display 1.2.1 produces the following output in the Session window.

```
Stem-and-leaf of time N = 66
Leaf Unit = 1.0
  1  -4 4
  1  -3
  1  -2
  1  -1
  2  -0 2
  2   0
  5   1 669
(41)  2 01122333444445555566666777777888888899999
 20   3 0001122222334666679
  1   4 0
```

It is a stem-and-leaf plot of the values in `time` with an increment of 10. Notice that we have placed 10 in the Increment box in the dialog box shown in Display 1.2.1 to reflect the fact we want the stem to be the units of 10.



Display 1.2.1: Dialog box for droducing a stem-and-leaf plot.

The first column gives the *depths* for a given stem, i.e., the number of observations on that line and below it or above it, depending on whether or not the observation is below or above the median. The row containing the median is enclosed in parentheses ( ), and the depth is only the observations on that line. If the number of observations is even and the median is the average of values on different rows, then parentheses do not appear. The second column gives the *stems,* as determined by what is placed in Increment, and the remaining columns give the ordered *leaves,* where each digit represents one observation. The *Leaf Unit* determines where the decimal place goes after each leaf. So in this example, the first observation is -44.0, while it would be -4.4 if the Leaf Unit were .1. Multiple stem-and-leaf plots can be carried out for a number of columns simultaneously and also for a single variable by the values of another variable.

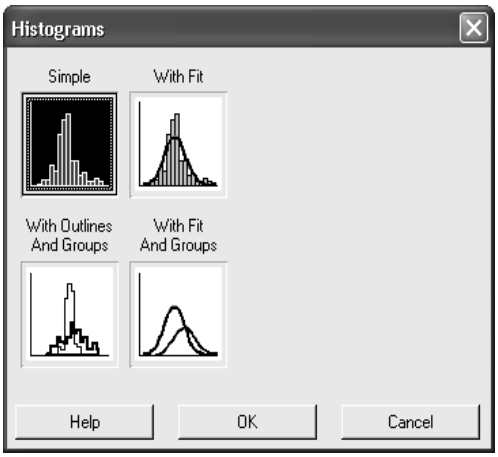
1.2.2 Histograms

A histogram is a plot where the data are grouped into intervals, and over each such interval a bar is drawn of height equal to the frequency (count) of data values in that interval (*frequency histogram*) or of height equal to the relative frequency (proportion) of data values in that interval (*relative frequency histogram*) or of height equal to the *density* of points in that interval, i.e., the proportion of points in the interval divided by the length of the interval (*density histogram*). We recommend plotting density histograms. The Graph ▶ Histogram command is used to obtain these plots.

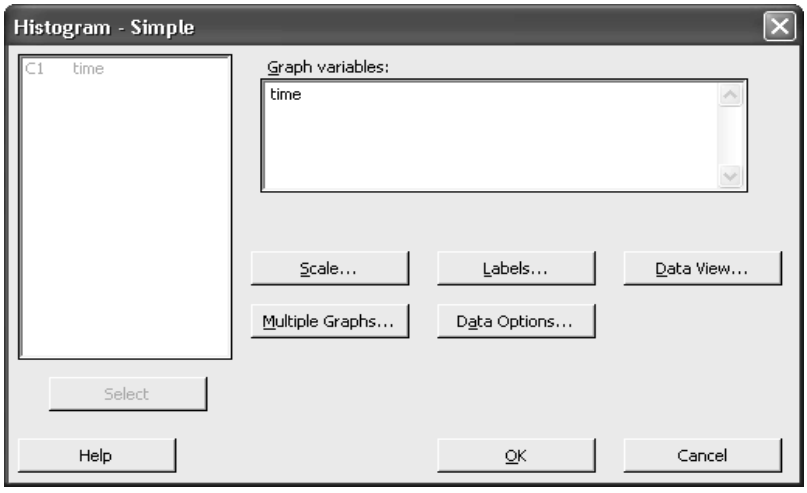
An important consideration when plotting multiple histograms for comparison purposes is to ensure that all the histograms have the same x and y scales so that the plots are visually comparable. The Graph ▶ Histogram command contains options that impose this restriction.

Using Graph ▶ Histogram with the `newcomb` worksheet, produces the dialog box shown in Display 1.2.2. Selecting Simple and clicking on OK leads to the dialog box in Display 1.2.3. We have placed the variable `time` in the Graph variables box to indicate we want a histogram of this variable. To select a density histogram we click on the Scale button, which brings up the dialog box of Display 1.2.4, and then click on the Y-scale Type to obtain the dialog box in Display 1.2.5, in which we have filled in the Density radio button. Clicking on OK in this dialog box and in the dialog box of Display 1.2.3 produces the density histogram of Display 1.2.6.

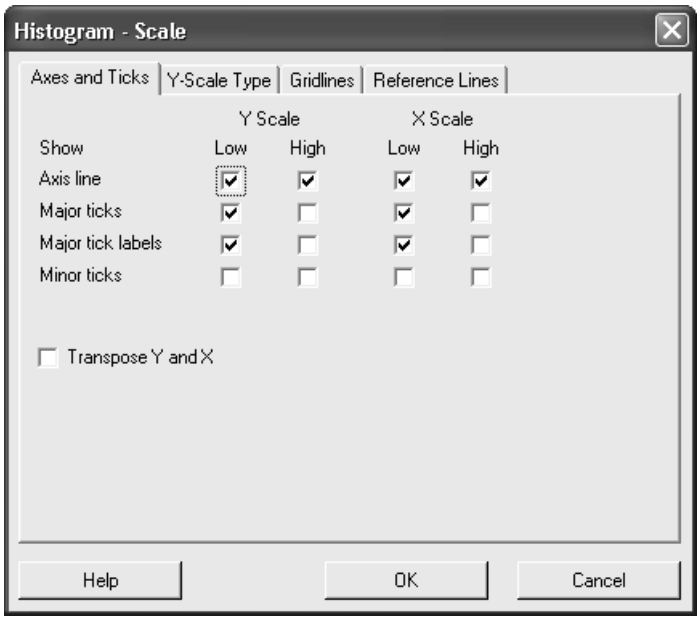
Note that we can produce multiple histograms by clicking on the Multiple Graphs button in the dialog box of Display 1.2.3.



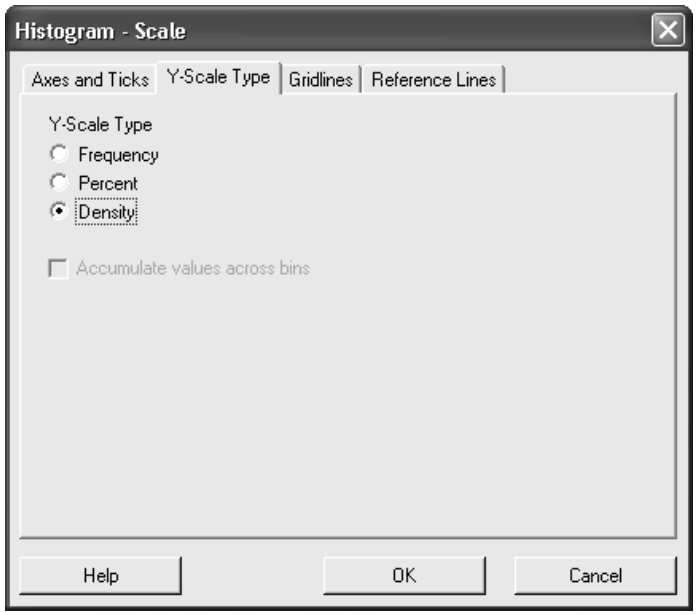
Display 1.2.2: Dialog box for selecting type of histogram.



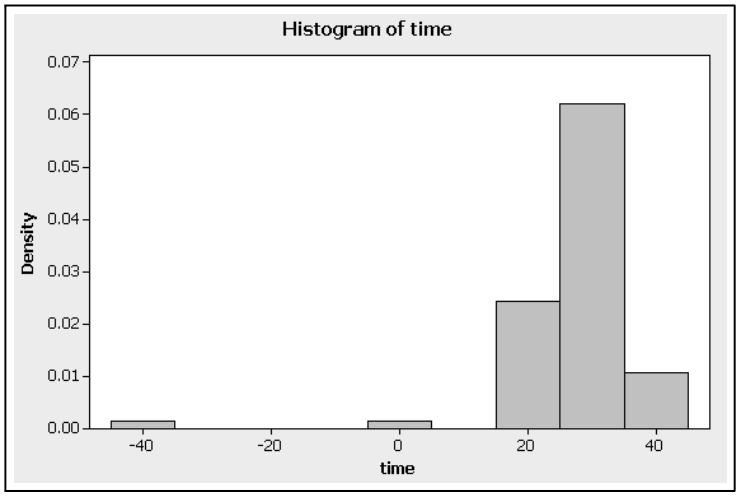
Display 1.2.3: Dialog box for creating a histogram of the `time` variable in the `newcomb` worksheet.



Display 1.2.4: Dialog box for specifying characteristics of the histogram plotted.

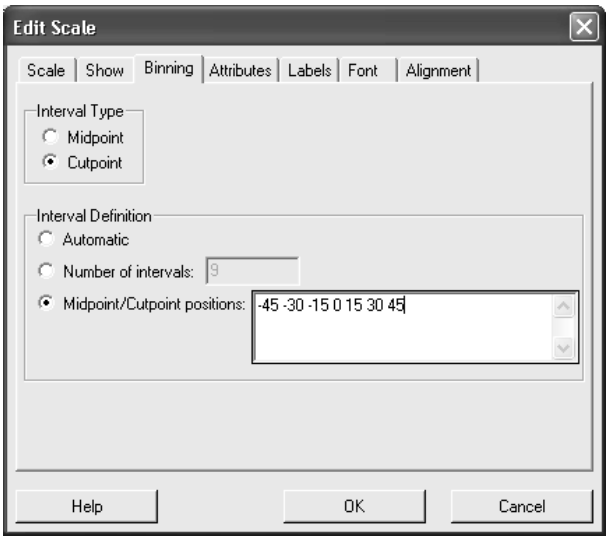


Display 1.2.5: Dialog box for selecting frequency, relative frequency or density histogram.

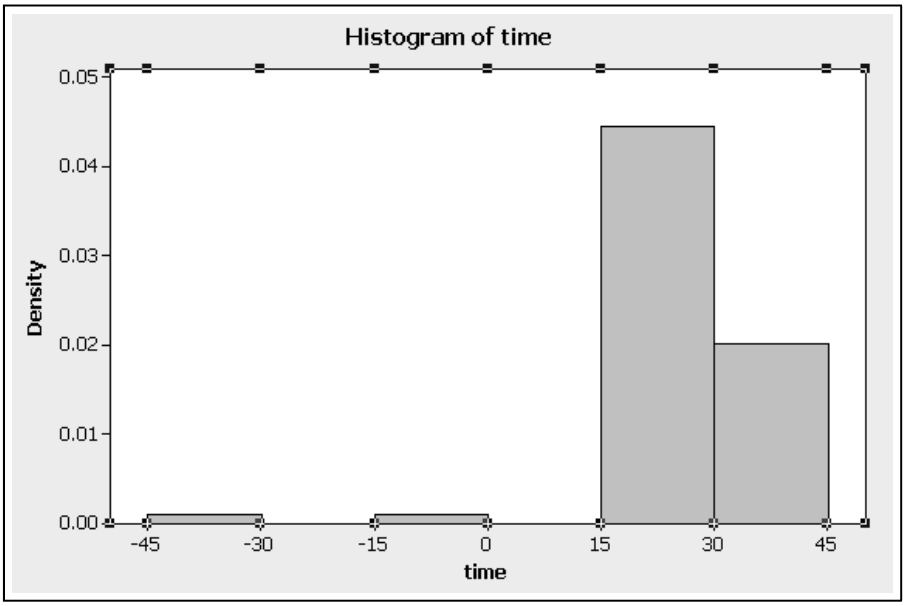


Display 1.2.6: Density histogram of the `time` variable in the `newcomb` worksheet.

We can also edit a graph to modify its appearance by double-clicking on various components of the plot in the graph window. For example, the plot in Display 1.2.6 is based on a default algorithm in Minitab to divide up the range of the data into bins and plot each bar over the mid-point of each bin. Sometimes we prefer to select the bins ourselves and moreover specify *cutpoints* (the end-points of each bin) rather than midpoints and have these cutpoints along the x-axis. To do this, we double click on a value on the x-axis which brings up the dialog box in Display 1.2.7, where we have clicked on the Binning tab. Here, we have selected the radio button Cutpoints in the Interval type box and have filled in the cutpoints $-45, -30, -15, 0, 15, 30, 45$ in the Midpoint/Cutpoint positions box. Clicking on OK produces the plot shown in Display 1.2.8.



Display 1.2.7: Dialog box for editing the bins for the histogram.



Display 1.2.8: Density histogram of the `time` variable in the `newcomb` worksheet with specified cutpoints.

The session command **histogram** is also available. This has the general syntax

histogram $E_1 \ldots E_m$

where E_1, ..., E_m correspond to columns. For example, the commands

```
MTB >histogram c1;
SUBC>cutpoints -45 -30 -15 0 15 30 45;
SUBC>density.
```

produce the histogram in Display 1.2.8 using the **cutpoints** and **density** subcommands. There are also subcommands **midpoints** and **nintervals,** which specify the number of subintervals, and **frequency** or **percent,** which respectively ensure that the heights of the bar lines equal the frequency and relative frequency of the data values in the interval. Also, the **cumulative** subcommand is available so that the bars represent all the values less than or equal to the endpoint of an interval. The subcommand **same** ensures that multiple histograms all have the same scale.

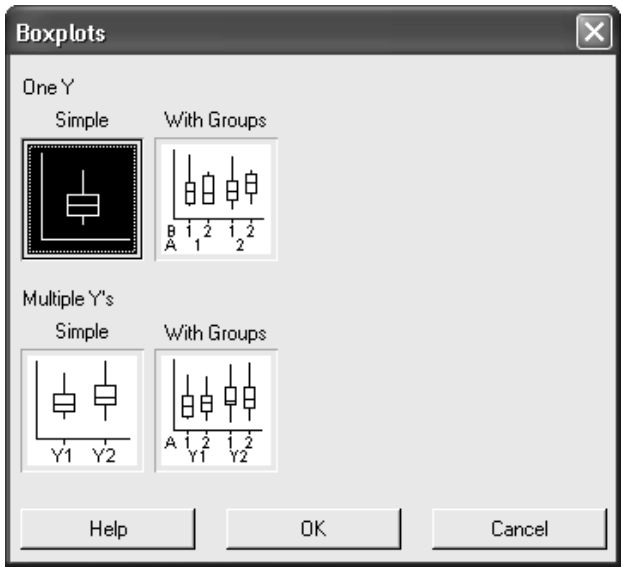
1.2.3 Boxplots

Boxplots are useful summaries of a quantitative variable and are obtained using the Graph ▶ Boxplot command. Boxplots are used to provide a graphical notion of the location of the data and its scatter in a concise and evocative way.

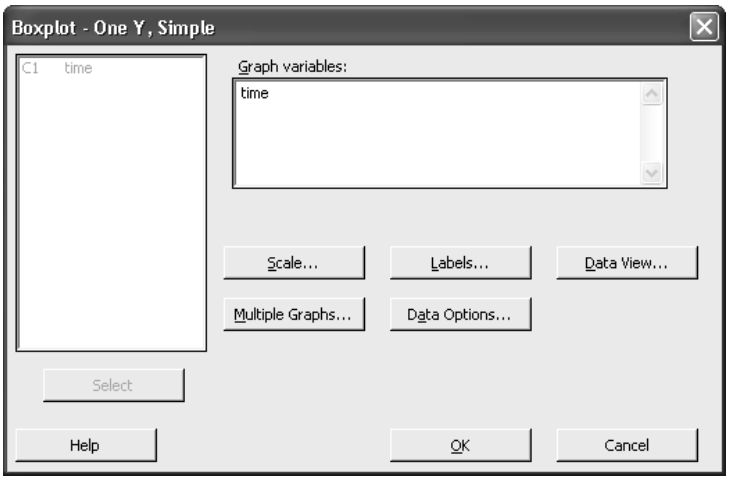
A boxplot is presented in Display 1.2.11 for the variable time in the `newcomb` worksheet. The line in the center of the box is the median. The line below the median is the first quartile, also called the *lower hinge,* and the line above is third quartile, also called the *upper hinge*. The difference between the third and first quartile is called the *interquartile range,* or IQR. The vertical lines from the hinges are called *whiskers,* and these run from the hinges to the *adjacent values.* The adjacent values are given by the greatest value less than or equal to the *upper limit* (the third quartile plus 1.5 times the IQR) and by the least value greater than or equal to the *lower limit* (the first quartile minus 1.5 times the IQR). The upper and lower limits are also referred to as the *inner fences.* The *outer fences* are defined by replacing the multiple 1.5 in the definition of the inner fences by 3.0. Values beyond the outer fences are plotted with a * and are called *outliers.* As with the plotting of histograms, multiple boxplots can be plotted for comparison purposes, and again, it is important to make sure that they all have the same scale.

The Graph ▶ Boxplot command produces the dialog box shown in Display 1.2.9. Selecting Simple and clicking on OK produces the dialog box shown in Display 1.2.10, where we have filled in the `time` variable in the Graph variable box. Clicking on OK produces the boxplot shown in Display 1.2.11.

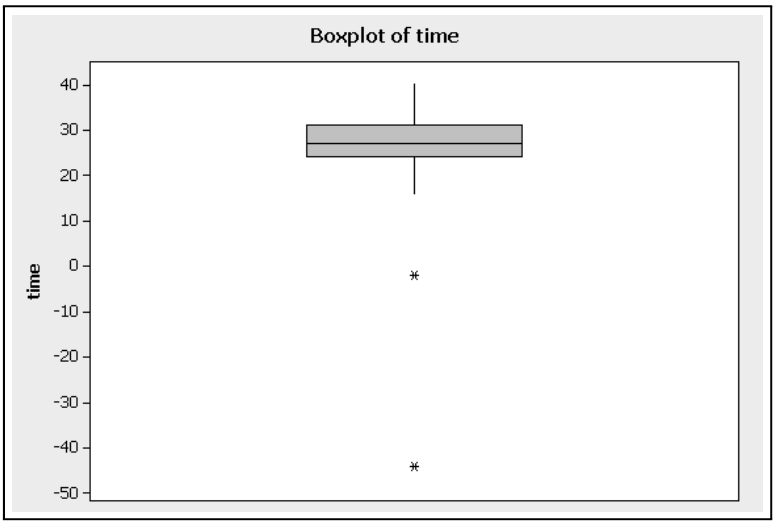
There is a corresponding session command called **boxplot.** We refer the reader to **help** for more discussion of this command.



Display 1.2.9: Dialog box for selecting type of boxplot.



Display 1.2.10: Dialog box for producing a boxplot of the `time` variable in the `newcomb` worksheet.



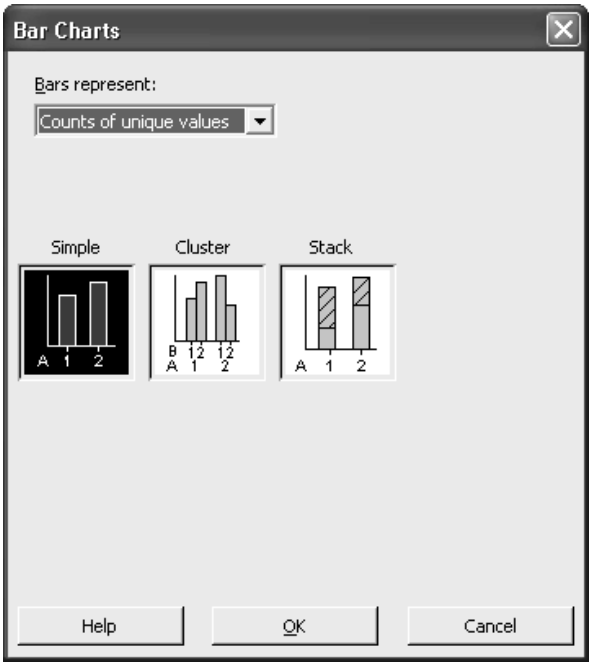
Display 1.2.11: Boxplot of the `time` variable in the `newcomb` worksheet.

There is a corresponding session command called **boxplot.** We refer the reader to **help** for more discussion of this command.

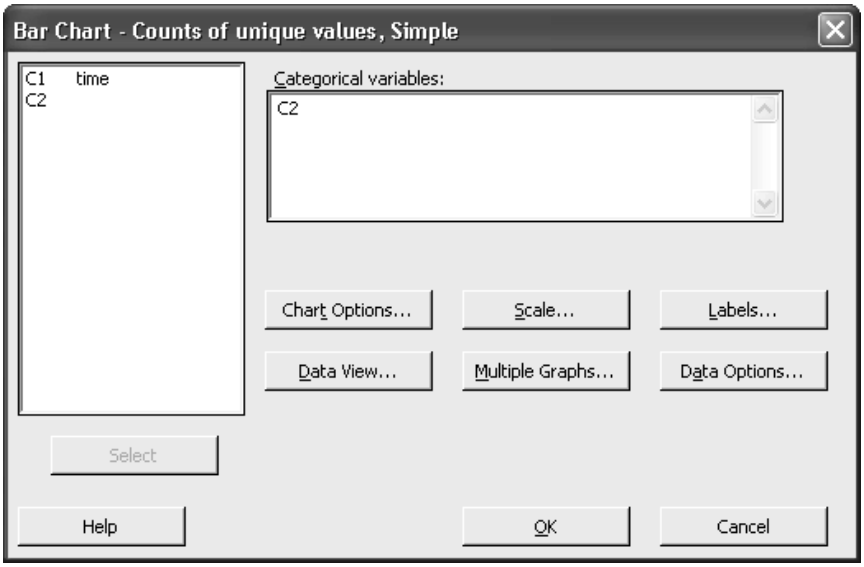
1.2.4 Bar Charts

Bar charts are used to plot the distributions of categorical variables. Consider the categorical variable C2 (created in Section 1.1.1) in the `newcomb` worksheet. The command Graph ▶ Bar Chart brings up the dialog box shown in Display 1.2.12. Selecting Simple and clicking on OK brings up the dialog box shown in Display 1.2.13, where we have filled in the Categorical variables box with C2. Now, since we want a graph of the distribution of C2, we next clicked on the

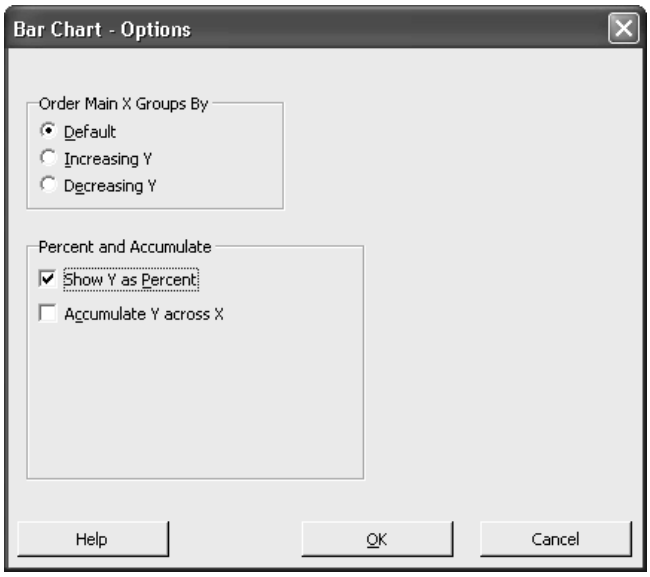
Chart Options button to bring up the dialog box of Display 1.2.14, where we have checked the Show Y as a Percent box. Clicking on OK in this and the dialog box of Display 1.2.13 produces the bar chart of Display 1.2.15.



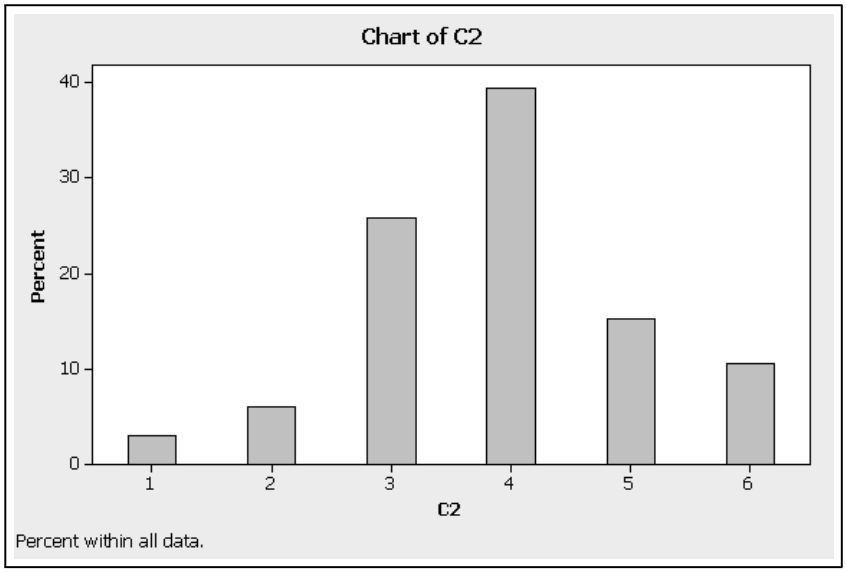
Display 1.2.12: Dialog box for selecting type of bar chart.



Display 1.2.13: Dialog box for selecting variable to plot in a bar chart.



Display 1.2.14: Dialog box to use to specify that you want the distribution to be plotted (and not just the counts).



Display 1.2.15: Bar chart of the variable C2 in the `newcomb` worksheet.

The corresponding session command is

chart E_1

which produces a bar chart for the values in column E_1. The subcommand **percent** ensures that the distribution is plotted.

1.2.5 Pie Charts

A *pie chart* is a disk divided up into wedges where each wedge corresponds to a unique value of a variable, and the area of the wedge is proportional to the relative frequency of the value with which it corresponds. Pie charts can be obtained via Graph ▶ Pie Chart, and there are various features available in the dialog box that can be used to enhance these plots. Pie charts are a common method for plotting categorical variables.

1.2.6 Time Series Plots

Often, data are collected sequentially in time. In such a context, it is instructive to plot the values of quantitative variables against time in a time series plot. For this we use the Graph ▶ Time Series Plot command. A discussion of these plots can be found in Chapter 18.

1.3 The Normal Distribution

It is important in statistics to be able to do computations with the normal distribution. The equation of the *density curve* for the normal distribution with mean μ and standard deviation σ is given by

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$$

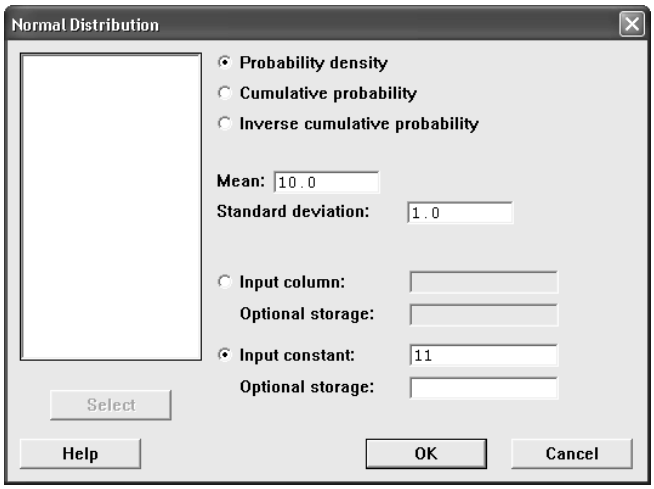
where z is a number. We refer to this as the $N(\mu, \sigma)$ (read as normal mu sigma) density curve or the $N(\mu, \sigma)$ density function. Note that notation for this function varies by text with many texts calling this the $N(\mu, \sigma^2)$ density function; i.e., the square of σ is used instead. So you have to be careful and check which notation your text is using.

Also of interest is the area under the density curve from $-\infty$ to a number x, i.e., the area between the graph of the $N(\mu, \sigma)$ density curve and the interval $(-\infty, x]$. This is a value between 0 and 1 and is referred to as the value of the $N(\mu, \sigma)$ distribution function at x.

Sometimes, we specify a value p between 0 and 1 and then want to find the point x_p, such that p of the area under the $N(\mu, \sigma)$ density curve lies over $(-\infty, x_p]$. The point x_p is called the *pth percentile* of the $N(\mu, \sigma)$ density curve. Alternatively, x_p is called the value of the inverse $N(\mu, \sigma)$ distribution function at p.

1.3.1 Calculating the Density

Suppose that we want to evaluate the $N(\mu, \sigma)$ density function at a value x. For this, we use the Calc ▶ Probability Distributions ▶ Normal command. For example, the dialog box in Display 1.3.1 calculates the $N(10, 1)$ density curve at the value $x = 11.0$.



Display 1.3.1: Dialog box for normal probability calculations.

After clicking on the OK button, the output

```
Normal with mean = 10 and standard deviation = 1
 x f( x )
11 0.241971
```

is printed in the Session window, which gives the value as 0.241971. Sometimes, we will want to evaluate the density curve at every value in a column of values, for example, when we are plotting this curve. For this, we simply click on the radio button Input column and type the relevant column in the associated box.

The general syntax of the corresponding session command **pdf** with the **normal** subcommand is

pdf $E_1 \ldots E_m$ into $E_{m+1} \ldots E_{2m}$;
normal mu = V_1 sigma = V_2.

where E_1, ..., E_m are columns or constants containing numbers and E_{m+1}, ..., E_{2m} are the columns or constants that store the values of the $N(\mu, \sigma)$ density curve at these numbers and $V_1 = \mu$ and $V_2 = \sigma$. If no storage is specified, then the values are printed. For example, if we want to compute the $N(-.5, 1.2)$ density curve at every value between -3 and 3 in increments of .01, the commands

```
MTB >set c1
DATA>-3:3/.01
DATA>end
MTB >pdf c1 c2;
SUBC>normal mu=-.5 sigma=1.2.
```

put the values between -3 and 3 in increments of .01 in C1 using the **set** command. The **pdf** command with the **normal** subcommand calculates the $N(-.5, 1.2)$ density curve at each of these values and puts the outcomes in the corresponding entries of C2. If we plot C2 against C1, we will have a plot of

the density curve of this distribution. For this, we use the scatterplot facilities in Minitab as discussed in II.2.1. Note that with the **normal** subcommand we must also specify the mean and the standard deviation via **mu** and **sigma.**

1.3.2 Calculating the Distribution Function

Suppose that we want to evaluate the area under $N(\mu, \sigma)$ density curve over the interval $(-\infty, x]$. This is the value of the cumulative distribution function of the $N(\mu, \sigma)$ distribution at the value x. For this, we use the Calc ▶ Probability Distributions ▶ Normal as well, but in this case, in the dialog box of Display 1.3.1, we select Cumulative probability instead. Making this change in the dialog box of Display 1.3.1, we get the output

```
Normal with mean = 10 and standard deviation = 1
 x P( X <= x )
11 0.841345
```

in the Session window. Again, we can evaluate this function at a single point or at every value in a variable.

The general syntax of the corresponding session command **cdf** command with the **normal** subcommand is

cdf $E_1 \ldots E_m$ into $E_{m+1} \ldots E_{2m}$;
normal mu = V_1 sigma = V_2.

where E_1, ..., E_m are columns or constants containing numbers and E_{m+1}, ..., E_{2m} are the columns or constants that store the values of the area under $N(\mu, \sigma)$ density curve over the interval from $-\infty$ to these numbers and $V_1 = \mu$ and $V_2 = \sigma$. If no storage is specified, the values are printed.

1.3.3 Calculating the Inverse Distribution Function

To evaluate inverse distribution function for the $N(\mu, \sigma)$ distribution, we again use the Calc ▶ Probability Distributions ▶ Normal command, but in this case, in the dialog box of Display 1.3.1, we select Inverse cumulative probability. Making this change in the dialog box of Display 1.3.1 and replacing 11 by .75—recall that the argument to this function must be between 0 and 1—we get the output

```
Normal with mean = 10 and standard deviation = 1
P( X <= x )     x
 0.75       10.6745
```

in the Session window. This indicates that the area to the left of 10.6745, underneath the $N(10, 1)$ density curve, is .75.

The general syntax of the corresponding session command **invcdf** with the **normal** subcommand is

invcdf $E_1 \ldots E_m$ into $E_{m+1} \ldots E_{2m}$;
normal mu = V_1 sigma = V_2.

where E_1, ..., E_m are columns or constants containing numbers between 0 and 1, and E_{m+1}, ..., E_{2m} are the columns or constants that store the values of the percentiles of the $N(\mu, \sigma)$ density curve at these numbers, and where $V_1 = \mu$ and $V_2 = \sigma$. If no storage is specified, then the values are printed.

1.3.4 Normal Probability Plots

Some statistical procedures require that we assume that values for some variables are a sample from a normal distribution. A *normal probability plot* checks for the reasonableness of this assumption. To create such a plot, we use the Graph ▶ Probability Plot command.

Suppose we want a normal probability plot for the `time` variable in the `newcomb worksheet.` Using Graph ▶ Probability Plot, we get the dialog box in Display 1.3.2, where we have selected Single and then clicked on OK. This brings up the dialog box in Display 1.3.3, where we placed `time` in the Variables box. Clicking on the Scale button and then the Y-Scale Type tab produces the dialog box of Display 1.3.4, where we have filled in the Scores option. Clicking on the OK button in this and the dialog box of Display 1.3.3 produces the plot in Display 1.3.5.

The normal probability plot is given by the symbol •. This plot should be like a straight line. It is not a straight line in this case and would appear to be clear evidence that the data do not come from a normal distribution. There are many other features available with these plots and we refer the reader to the online manual for a discussion of these.

It should be noted that Minitab computes the (normal) scores as follows. For an observation that has rank i, the normal score is calculated as

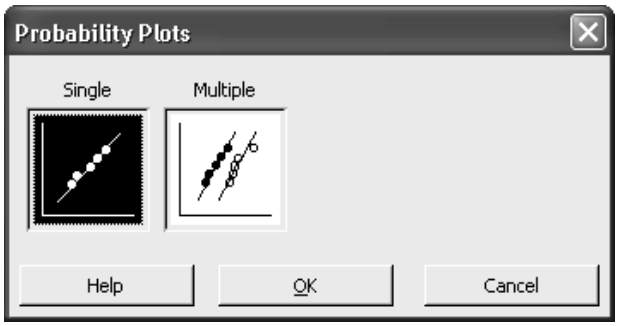
$$\Phi^{-1}\left((i - .375) \,/\, (n + .25)\right).$$

In Display 1.3.4, the values $(i - .375) \,/\, (n + .25)$ are referred to as probabilities, while $100\,(i - .375) \,/\, (n + .25)$ are referred to as percents.

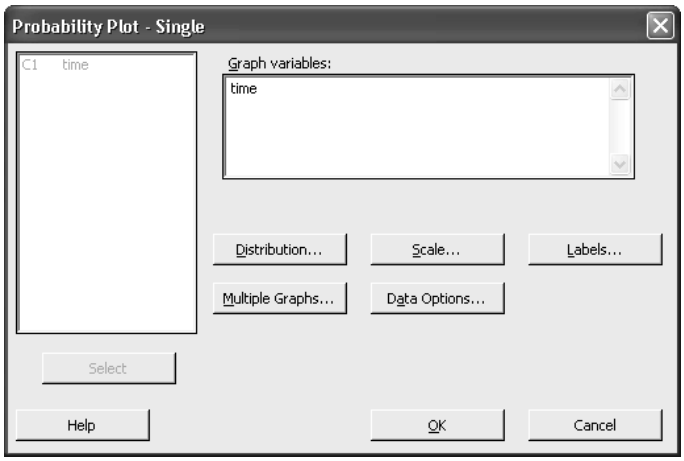
The session commands

```
MTB >nscores c1 c3
MTB >plot c3*c1
```

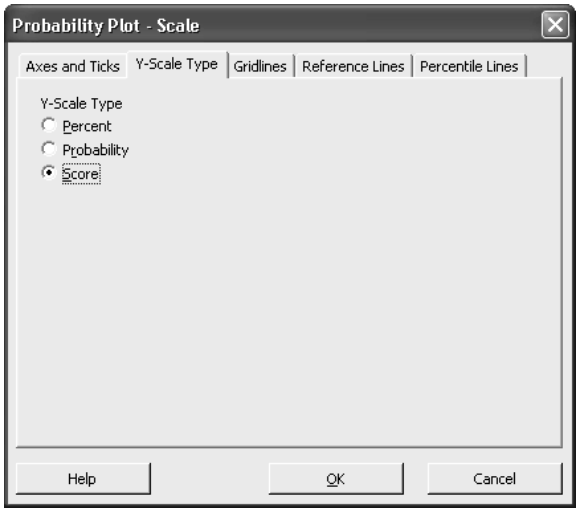
produce a normal probability plot like that shown in Display 1.3.5. The **nscores** (*normal scores*) command computes the score for each observation in C1 and places this in the corresponding entry of C3. The **plot** command then plots C3 versus C1 in a scatterplot.



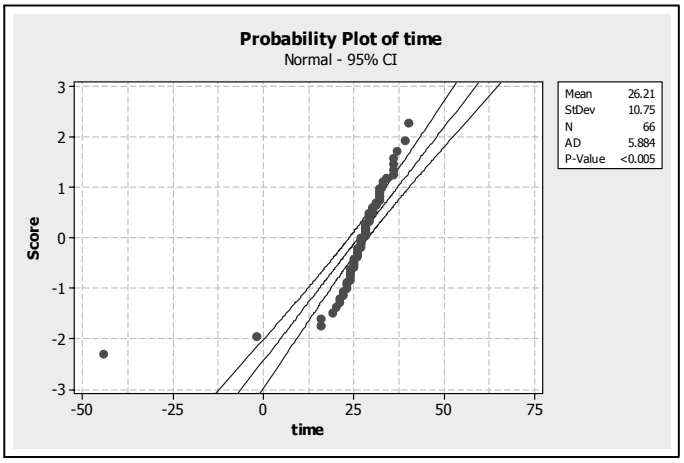
Display 1.3.2: First dialog box for producing a normal probability plot.



Display 1.3.3: Second dialog box for producing normal probability plots.



Display 1.3.4: Dialog box for selecting the Y-scale in a normal probability plot.



Display 1.3.5: Normal probability plot for the `time` variable in the `newcomb` worksheet.

1.4 Exercises

1. Using Newcomb's measurements in Table 1.1.1, create a new variable by grouping these values into three subintervals $[-50, 0), [0, 20), [20, 50)$. Calculate the frequency distribution, the relative frequency distribution, and the cumulative distribution of this ordered categorical variable.

2. Using Newcomb's measurements in Table 1.1.1, calculate and print the empirical distribution function. From this, determine the first quartile, median, and third quartile. Also, use the empirical distribution function to compute the 10th and 90th percentiles.

3. Consider the following sample of $n = 20$ data values.

1.3	0.7	0.7	−1.0	2.5	−0.1	−0.2	−0.1	1.7	0.0
1.1	−1.1	2.1	−0.9	−0.3	−1.0	−2.4	−0.6	−0.3	3.3

   Produce a stemplot of these data and determine the median.

4. For the data in Exercise 1.3 use an appropriate Minitab command to determine the minimum, maximum, first and third quartiles and median.

5. Transform the data in Exercise 1.3 by adding 3 to each data value and repeat Exercise 1.4. What do you notice?

6. Transform the data in Exercise 1.3 by subtracting 5 from each value and multiplying by 10. Calculate the means and standard deviations, using any Minitab commands, of both the original and transformed data. Compute the ratio of the standard deviation of the transformed data to the standard deviation of the original data. Comment on this value.

7. Transform the data in Exercise 1.3 by multiplying each value by 3. Compute the ratio of the standard deviation to the mean (called the *coefficient of variation*) for the original data and for the transformed data. Justify the outcome.

8. For the $N(6, 1.1)$ density curve, compute the area between the interval $(3, 5)$ and the density curve. What number has 53% of the area to the left of it for this density curve?

9. Use Minitab commands to verify the 68-95-99.7 rule for the $N(2, 3)$ density curve.

10. Calculate and store the values of the $N(0, 1)$ density curve at each value in $[-3, 3]$ using an increment of .01. Put the values in the interval $[-3, 3]$ in C1 and the values of the density curve in C2. Using the command `plot C2*C1`, plot the density curve. Comment on the shape of this curve.

11. For the data in Exercise 1.3 produce a normal quantile plot and comment on the validity of assuming that this is a sample from a normal distribution.

Chapter 2

Looking at Data: Exploring Relationships

New Minitab commands discussed in this chapter

Graph ▶ Plot
Stat ▶ Basic Statistics ▶ Correlation
Stat ▶ Regression ▶ Fitted Line Plot
Stat ▶ Regression ▶ Regression

In this chapter, Minitab commands that permit the analysis of relationships among two variables are described. The methods are different depending on whether both variables are quantitative, both variables are categorical, or one is quantitative and the other is categorical. This chapter considers relationships between two quantitative variables with the remaining cases discussed in later chapters. Graphical methods are very useful in looking for relationships among variables, and we examine various plots for this.

2.1 Scatterplots

A *scatterplot* of two quantitative variables is a useful technique when looking for a relationship between two variables. By a scatterplot we mean a plot of one variable on the y-axis against the other variable on the x-axis.

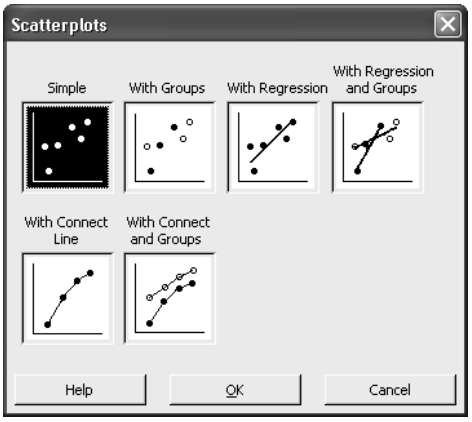
For example, consider the data in Table 2.1.1 collected from five fossil specimens of the extinct bird Archaeopteryx, where `femur` is the length in centimeters of the femur and `humerus` is the length in centimeters of the humerus. Here we are concerned with the relationship between the length of the femur and the length of the humerus. Suppose that we have input the data so that length of the humerus measurements are in C1, which has been named `humerus`, and the

length of the femur measurements are in C2, which has been named `femur`, of the worksheet `archaeopteryx`.

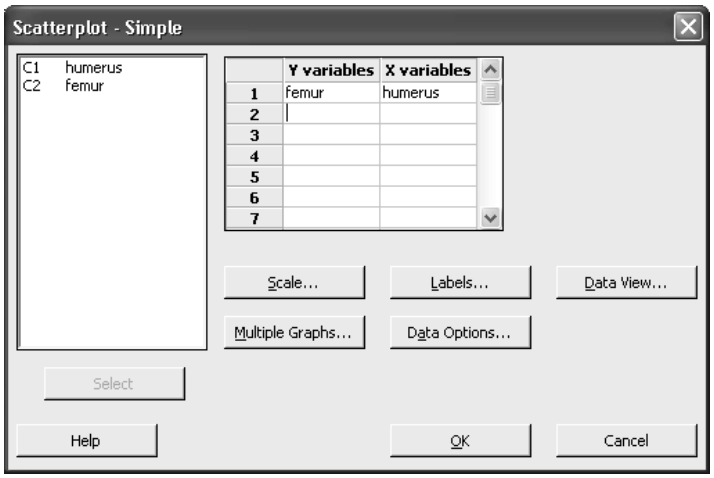
humerus	38	56	59	64	74
femur	41	63	70	72	82

Table 2.1.1: Archaeopteryx data.

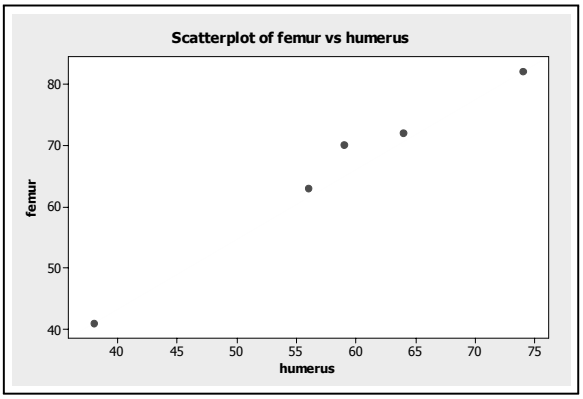
To plot the values of C2 against C1, we apply the Graph ▶ Scatterplot command to the contents of C1 and C2. First, we obtain Display 2.1.1 and from this we select Simple and click OK, which leads to the dialog box in Display 2.1.2. We then fill in C2 for the Y variable and C1 for the X variable. The plot depicted in Display 2.1.3 is produced in a separate Graph window when we click on OK.



Display 2.1.1: Dialog box for selecting the columns in a scatterplot.



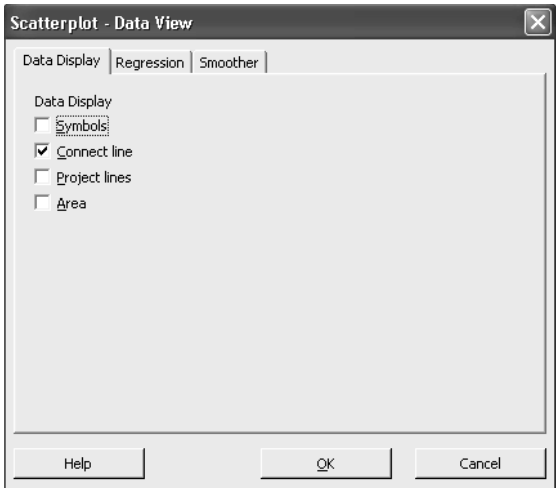
Display 2.1.2: Dialog box for selecting the columns in a scatterplot.



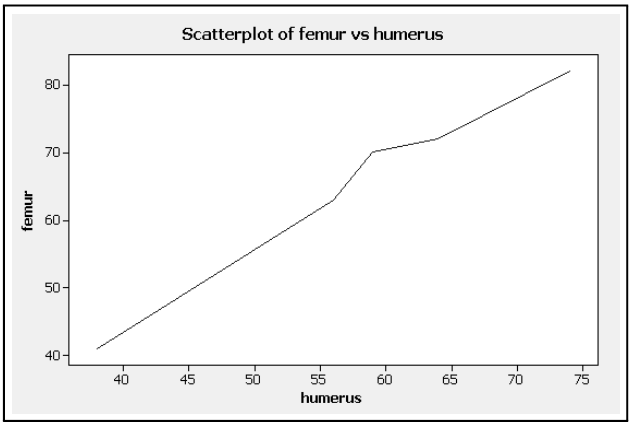
Display 2.1.3: Scatterplot of femur length (C2) versus humerus length (C1).

Note that the plotting symbol used in Display 2.1.3 for each point (x, y) is $\bullet$. Alternatives are available. Clicking on Data View in the dialog box of Display 2.1.2 leads to the dialog box of Display 2.1.4. If we select Connect line and plot the graph, we obtain the plot shown in Display 2.1.5. Also, you can add *projection lines* (drop a line from each point to the x-axis), and add *areas* (fill in the area under a polygon joining the points). Furthermore, you can employ the scatterplot smoother *lowess* to plot a piecewise linear continuous curve through the scatter of points (look under Smoother). The plot itself can be edited by clicking on objects in the plot.

There are a number of other features that allow you to control the appearance of the plot. In particular, you can double click any element of the plot and possibly modify its appearance according to the selections offered in the drop-down list that appears. For example, if we double click the plotted curve, we have the option of changing the plotting symbol and its size. We refer the reader to the online manual for a full description of this feature.



Display 2.1.4: Dialog box for selecting the appearance of the plotted line.



Display 2.1.5: Scatterplot with connecting lines.

The corresponding session command is **plot.** For example,

```
MTB > plot femur*humerus
```

produces a plot like that shown in Display 2.1.3. Note that the first variable is plotted along the y-axis, and the second variable is plotted along the x-axis. There are various subcommands that can be used with **plot,** and we refer the reader to Help for a description of these.

2.2 Correlations

While a scatterplot is a convenient graphical method for assessing whether or not there is any relationship between two variables, we would also like to assess this numerically. The *correlation coefficient* provides a numerical summarization of the degree to which a linear relationship exists between two quantitative variables, and this can be calculated using the Stat ▶ Basic Statistics ▶ Correlation command. For example, applying this command to the `femur` and `humerus` variables of the worksheet `archaeopteryx`, i.e., the data in Table 2.1.1 and depicted in Display 2.1.3, we obtain the output

```
Pearson correlation of humerus and femur = 0.991
P-Value = 0.001
```

in the Session window. For now, we ignore the number recorded as `P-Value`.

The general syntax of the corresponding session command **correlate** is given by

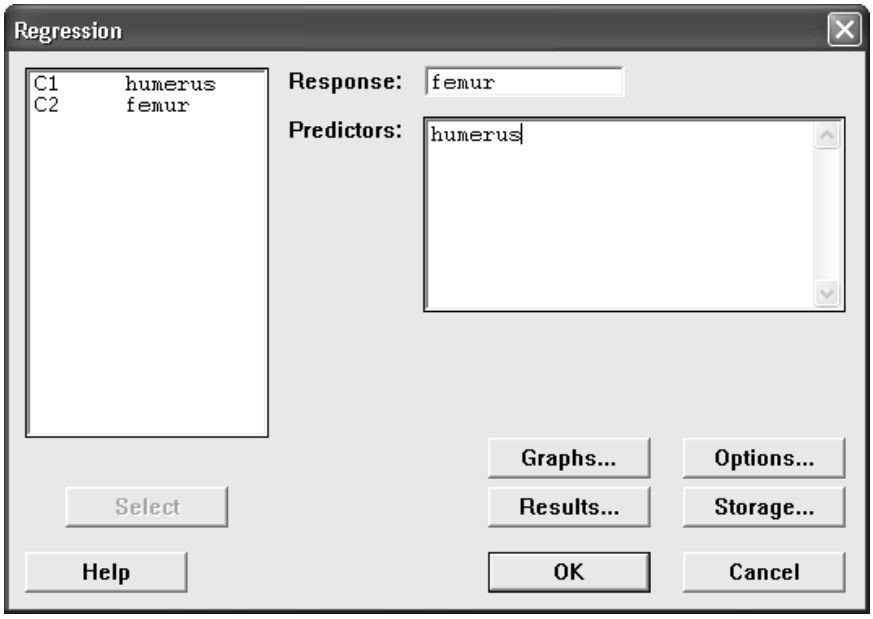
correlate $\mathrm{E}_1 \ldots \mathrm{E}_m$

where E_1, ..., E_m are columns corresponding to numerical variables, and a correlation coefficient is computed between each pair. This gives $m(m-1)/2$ correlation coefficients. The subcommand **nopvalues** is available if you want to suppress the printing of P-values.

2.3 Regression

Regression is another technique for assessing the strength of a linear relationship existing between two variables and it is closely related to correlation. For this, we use the Stat ▶ Regression command.

A regression analysis of two quantitative variables involves computing the least-squares line $y = a + bx$, where one variable is taken to be the response variable y and the other is taken to be the explanatory or predictor variable x. Note that the least-squares line is different depending upon which choice is made. For example, for the data of the worksheet `archaeopteryx`, i.e., the data in Table 2.1.1 and depicted in Display 2.1.3, letting `femur` be the response and `humerus` be the predictor or explanatory variable, the Stat ▶ Regression ▶ Regression command leads to the dialog box of Display 2.3.1, where we have made the appropriate entries in the Response and Predictors boxes. Clicking on the OK button leads to the output of Display 2.3.2 being printed in the Session window. This gives the least-squares line as $y = -1.42 + 1.15x$, i.e., $a = -1.420$ and $b = 1.15155$, which we also see under the `Coef` column in the first table. In addition, we obtain the value of the square of the correlation coefficient, also known as the *coefficient of determination,* as `R-Sq = 98.2%`. We will discuss the remaining output from this command in Chapter 10.



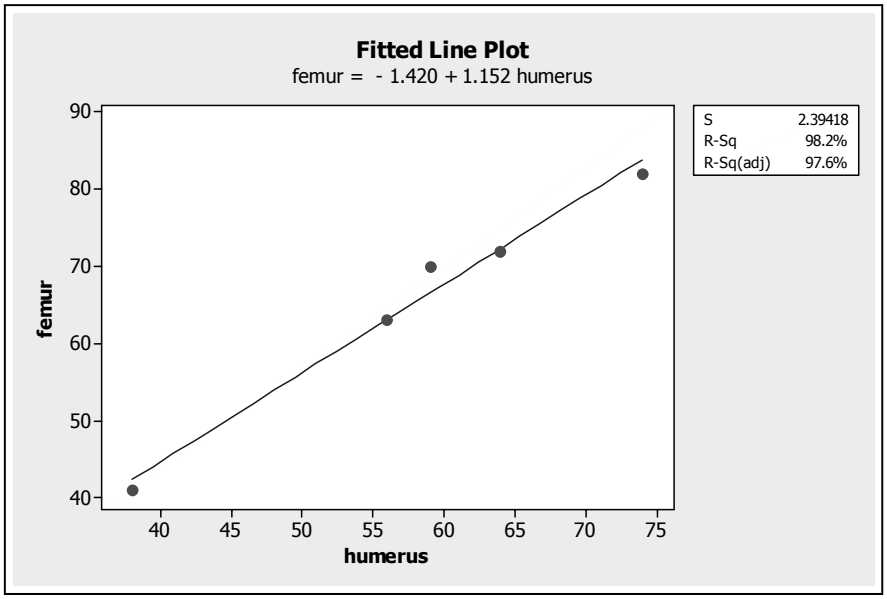
Display 2.3.1: Dialog box for a regression analysis.

Regression Analysis: femur versus humerus

```
The regression equation is
femur = - 1.42 + 1.15 humerus


Predictor     Coef  SE Coef      T      P
Constant    -1.420    5.386  -0.26  0.809
humerus    1.15155  0.09070  12.70  0.001


S = 2.39418   R-Sq = 98.2%   R-Sq(adj) = 97.6%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  924.00  924.00  161.20  0.001
Residual Error   3   17.20    5.73
Total            4  941.20
```

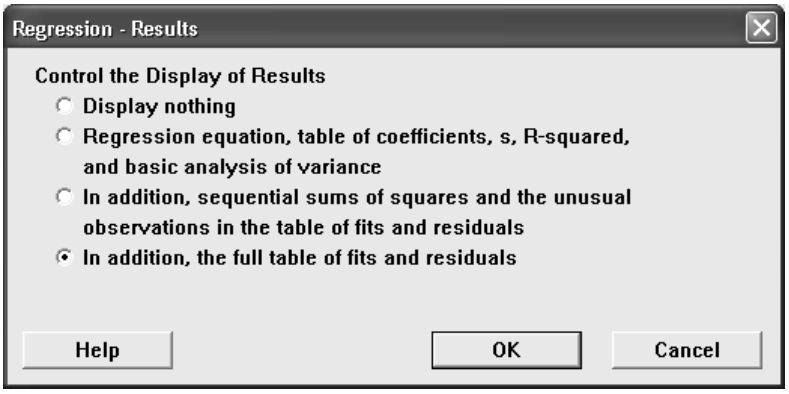
Display 2.3.2: Output from the dialog box of Display 2.3.1.

It is very convenient to have a scatterplot of the points together with the least-squares line. This can be accomplished using the Stat ▶ Regression ▶ Fitted Line Plot command. Filling in the dialog box for this command as in Display 2.3.1 produces the output in the Session window of Display 2.3.2 together with the plot of Display 2.3.3.



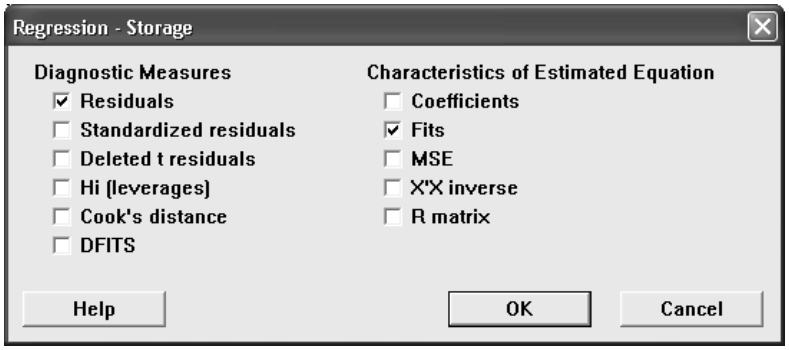
Display 2.3.3: Scatterplot of `femur` versus `humerus` in the `archaeopteryx` worksheet together with the least-squares line.

There are some additional quantities that are often of interest in a regression analysis. For example, you may wish to have the fitted values $\hat{y} = a + bx$ at each x value printed as well as the residuals $y - \hat{y}$. Clicking on the Results button in the dialog box of Display 2.3.1 and filling in the ensuing dialog box as in Display 2.3.4 results in these quantities being printed in the Session window as well as the output of Display 2.3.2.



Display 2.3.4: Dialog box for controlling output for a regression.

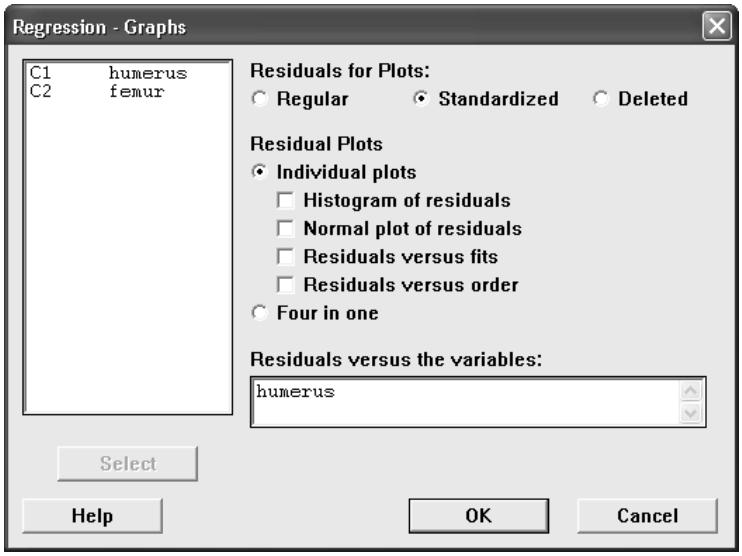
You will probably want to keep these values for later work. In this case, clicking on the Storage button of Display 2.3.1 and filling in the ensuing dialog box as in Display 2.3.5 results in these quantities being saved in the next two available columns—in this case, C3 and C4—with the names `resl1` and `fits1` for the residuals and fits, respectively.



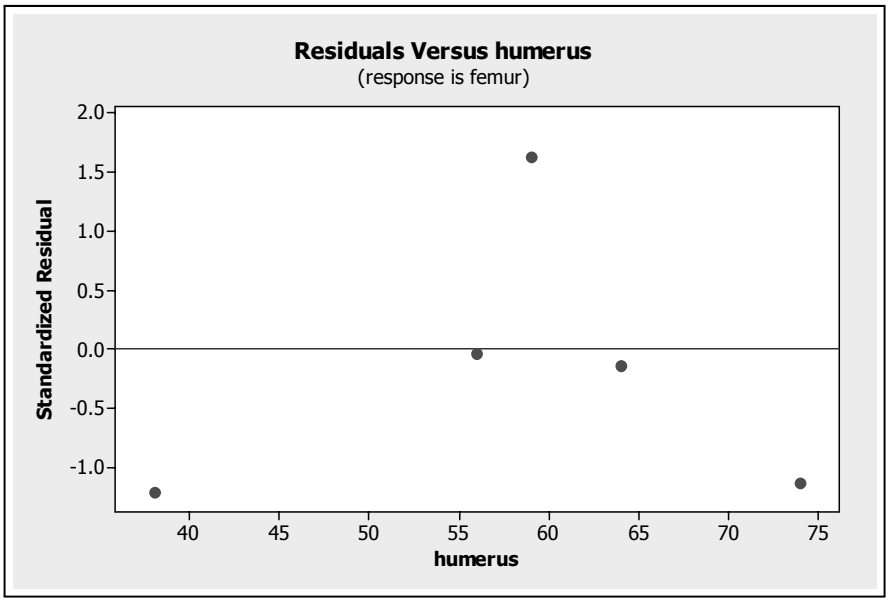
Display 2.3.5: Dialog box for storing various quantities computed in a regression.

Even more likely is that you will want to plot the residuals as part of assessing whether or not the assumptions that underlie a regression analysis make sense in the particular application. For this, click on the Graphs button in the dialog box of Display 2.3.1. The dialog box of Display 2.3.6 becomes available. Notice that we have requested that the *standardized residuals*—each residual divided by its standard error—be plotted, and this plot appears in Display 2.3.7. All the standardized residuals should be in the interval $(-3, 3)$, and no pattern should

be discernible. In this case, this residual plot looks fine. From the dialog box of Display 2.3.6, we see that there are many other possibilities for residual plots.



Display 2.3.6: Dialog box for selecting various residual plots as part of a regression.



Display 2.3.7: Plot of the standardized residuals versus `humerus` after regressing `femur` against `humerus` in the `archaeopteryx` worksheet.

The corresponding session command is given by **regress,** and by using the subcommands **pfits, residual,** and **sresidual** we can calculate and store *fitted values, residuals,* and *standardized residuals,* respectively. For example,

```
MTB > regress c1 1 c2;
SUBC> fits c3;
SUBC> residuals c4;
SUBC> sresiduals c5.
```

gives the output of Display 2.3.2 and also stores the fitted values in C3, stores the residuals $y - \hat{y}$ in C4, and stores the standardized residuals in C5. Note that the 1 in `regress c1 1 c2` refers to the number of predictors we are using to predict the response variable. To plot the standardized residuals against `humerus`, we use

```
MTB > plot c5*c2
```

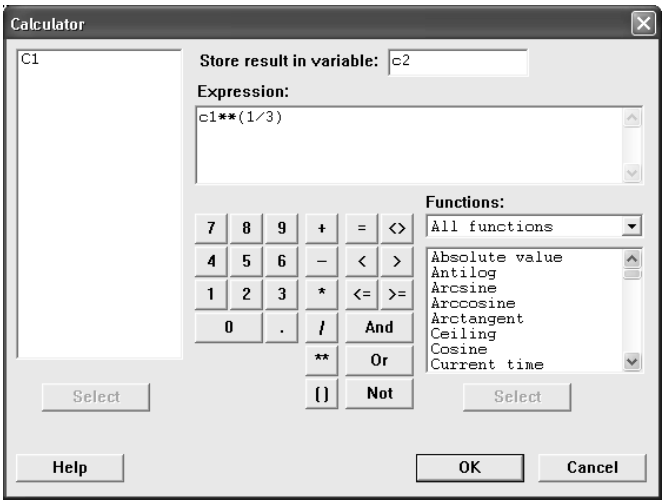
which results in a plot like Display 2.3.7 but with different labels on the x axis.

2.4 Transformations

Sometimes, transformations of the variables are appropriate before we carry out a regression analysis. This is accomplished in Minitab using the Calc ▶ Calculator command and the arithmetical and mathematical operations discussed in Sections I.10.1 and I.10.2. In particular, when a residual plot looks bad, sometimes this can be fixed by transforming one or more of the variables using a simple transformation, such as replacing the response variable by its logarithm or something else. For example, if we want to calculate the cube root—i.e., $x^{1/3}$—of every value in C1 and place these in C2, we use the Calc ▶ Calculator command and the dialog box as depicted in Display 2.4.1. Alternatively, we could use the session command **let** as in

```
MTB > let c2=c1**(1/3)
```

to produce the same result.



Display 2.4.1: Dialog box for calculating transformations of variables.

2.5 Exercises

1. Suppose the following data has been collected for two variables x and y, where y is the response and x is the predictor.

x	-0.5	-2.3	1.8	-3.0	2.1	-3.3	1.0	1.3	-1.9
y	0.6	-3.4	7.4	-3.8	7.8	-3.5	3.3	4.9	0.2

   Calculate the least-squares line and make a scatterplot of y against x together with the least-squares line. Plot the standardized residuals against x. What is the squared correlation coefficient between these variables?

2. Suppose in Exercise 2.5.1 there is another variable z where z takes the value 1 for the first five (x, y) pairs and takes the value 2 for the last four (x, y) pairs. Make a scatterplot of y against x where the points for different z are labeled differently (use Minitab for the labeling, too) and with the least-squares line on it.

3. Place the values 1 through 100 with an increment of .1 in C1 and the square of these values in C2. Calculate the correlation coefficient between C1 and C2. Multiply each value in C1 by 10, add 5, and place the results in C3. Calculate the correlation coefficient between C2 and C3. Why are these correlation coefficients the same?

4. Place the values 1 through 100 with an increment of .1 in C1 and the square of these values in C2. Calculate the least-squares line with C2 as response and C1 as explanatory variable. Plot the standardized residuals. If you see such a pattern of residuals, what transformation might you use to remedy the problem?

5. For the data in Exercise 2.5.1, numerically verify the algebraic relationship that exists between the correlation coefficient and the slope of the least-squares line using Minitab commands.

6. For the data in Exercise 2.5.1, calculate the sum of the residuals and the sum of the squared residuals divided by the number of data points minus 2. Is there anything you can say about what these quantities are equal to in general?

7. Place the values 1 through 10 with an increment of .1 in C1, and place x^3 of these values in C2. Calculate the least-squares line using C2 as the response variable, and plot the standardized residuals against C1. What transformation would you use to remedy this residual plot? What is the least-squares line when you carry out this transformation?

8. Place the values 1 through 10 with an increment of .1 in C1, and place $\exp(-1 + 2x)$ of these values in C2. Calculate the least-squares line using C2 as the response variable, and plot the standardized residuals against C1. What transformation would you use to remedy this residual plot? What is the least-squares line when you carry out this transformation?

Chapter 3

Producing Data

New Minitab commands discussed in this chapter

Calc ▶ Set Base
Calc ▶ Random Data

This chapter is concerned with the collection of data, perhaps the most important step in a statistical problem, as this determines the quality of whatever conclusions are subsequently drawn. A poor analysis can be fixed if the data are collected correctly by simply redoing the analysis. But if the data have not been appropriately collected, then no amount of analysis can rescue the study. We discuss Minitab commands that enable you to generate samples from populations and also to randomly allocate treatments to experimental units.

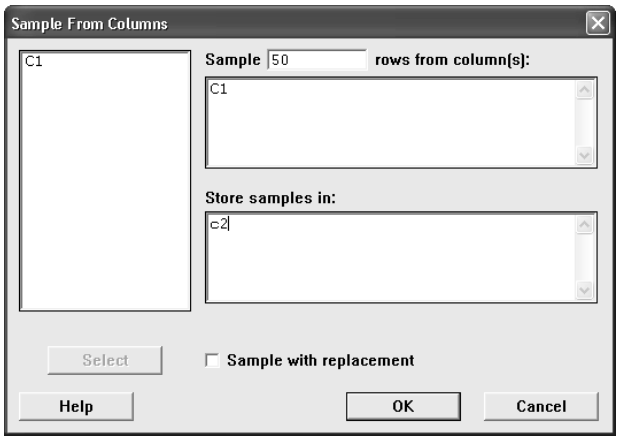
Minitab uses computer algorithms to mimic randomness. Still, the results are not truly random. In fact, any simulation in Minitab can be repeated, with exactly the same results being obtained, using the Calc ▶ Set Base command. For example, in the dialog box of Display 3.1, we have specified the base, or seed, random number as 1111089. The base can be any integer. When you want to repeat the simulation, you give this command, with the same integer. Provided you use the same simulation commands, you will get the same results. This can also be accomplished using the session command **base** `V`, where `V` is an integer.

Display 3.1: Dialog box for setting base or seed random number.

3.1 Generating a Random Sample

Suppose that we have a large population of size N and we want to select a sample of $n < N$ from the population. Further, we suppose that the elements of the population are ordered; i.e., we have been able to assign a unique number $1, \ldots, N$ to each element of the population. To avoid selection biases, we want this to be a *random sample;* i.e., every subset of size n from the population has the same "chance" of being selected. This implies that we generate our sample so that every subset of size n in the population has the same chance of being chosen. We can do this physically by using some simple random system, such as chips in a bowl or coin tossing. We could also use a table of random numbers, or, more conveniently, we can use computer algorithms that mimic the behavior of random systems.

For example, suppose there are 1000 elements in a population, and we want to generate a sample of 50 from this population without replacement. We can use the Calc ▶ Random Data ▶ Sample from Columns command to do this. For example, suppose we have labeled each element of the population with a unique number in $1, 2, \ldots, 1000$, and, further, we have put these numbers in C1 of a worksheet. The dialog box of Display 3.1.1 results in a random sample of 50 being generated without replacement from C1 and stored in C2.



Display 3.1.1: Dialog box for generating a random sample without replacement.

Printing this sample gives the output

```
MTB > print c2
C2
 211 609 690 869 257 145 700 756 830 864 953 155 747
 238 271 557 740 551 249 450 167 900 702 599 555  85
 926 933 628  21 880 191 189 750 804 991  47  53 202
 918 188 479 118 988 244 644 878 729 353 411
```

in the Session window. So now we go to the population and select the elements labeled 211, 609, 690, etc. The algorithm that underlies this command is such that we can be confident that this sample of 50 is like a random sample.

Sometimes we want to generate *random permutations,* i.e., $n = N$, and we are simply reordering the elements of the population. For example, in experimental design, suppose we have $N = n_1 + \cdots + n_k$ experimental units and k treatments, and we want to allocate n_i applications of treatment i. Suppose further that we want all possible such applications to be equally likely. Then we generate a random permutation $(l_1, \ldots, l_N)$ of $(1, \ldots, N)$ and allocate treatment 1 to those experimental units labeled $l_1, \ldots, l_{n_1}$, allocate treatment 2 to those experimental units labeled $l_{n_1+1}, \ldots, l_{n_1+n_2}$, etc. For example, if we have 30 experimental units and 3 treatments and we want to allocate 10 experimental units to each treatment, placing the numbers $1, 2, \ldots, 30$ in C1 and using the Calc ▶ Random Data ▶ Sample from Columns command as in the dialog box of Display 3.1.1, but with 30 in the Sample box, generates a random permutation of $1, 2, \ldots, 30$ in C2. Implementing this gives us the random permutation

```
MTB > print c2
C2
 13  7 26  8 22 23 28 17  3 25
 9   2 14 29 15 18  6 11 16  5
 12 27  4 30 20 24  1 19 21 10
```

and for the treatment allocation you can read the numbers row-wise or column-wise, as long as you are consistent. Row-wise is probably best, as this is how the numbers are stored in C2, and so you can always refer back to C2 (presuming you save your worksheet) if you get mixed up.

The above examples show how to directly generate a sample from a population of modest size. But what happens if the population is huge or it is not convenient to label each unit with a number? For example, suppose we have a population of size 100,000 for which we have an ordered list and we want a sample of size 100. In this case, more sophisticated techniques need to be used, but simple random sampling can still typically be accomplished (see Exercise 3.3 for a simple method that works in some contexts).

Simple random sampling corresponds to sampling without replacement; i.e., after we randomly select an element from the population, we do not return it to the population before selecting the next sample element. Sampling with replacement corresponds to replacing each sample element in the population after selecting it and recording only the element that was obtained. So at each selection, every element has the same chance of being selected, and an element may appear more than once in the sample. Notice that we can also sample with replacement if we check the Sample with replacement box in the dialog box of Display 3.1.1.

The general syntax of the corresponding session command **sample** is

sample V $\mathrm{E}_1 \ldots \mathrm{E}_m$ put into $\mathrm{E}_{m+1} \ldots \mathrm{E}_{2m}$

where V is the sample size n and V rows are sampled from the columns E_1, ..., E_m and stored in columns E_{m+1}, ..., E_{2m}. If we wanted to sample with replacement—i.e., after a unit is sampled, it is placed back in the population so that it can possibly be sampled again—we use the **replace** subcommand. Of

course, for simple random sampling, we do not use the **replace** subcommand. Note that the columns can be numeric or text.

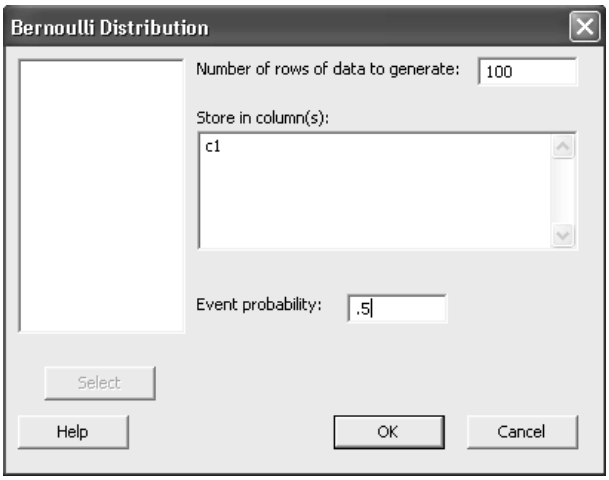
3.2 Sampling from Distributions

Once we have generated a sample from a population, we measure various attributes of the sampled elements. For example, if we were sampling from a population of humans, we might measure each sampled unit's height. The height for the sample unit is now a random variable that follows the height distribution in the population from which we are sampling. For example, if 80% of the people in the population are between 4.5 feet and 6 feet, then under *repeated sampling* of an element from the population (with replacement) in the long run, 80% of the sampled units will have their heights in this range.

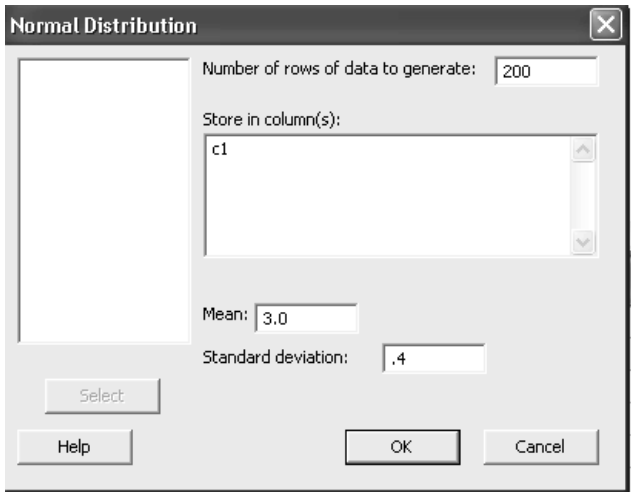
Sometimes, we want to sample directly from this population distribution; i.e., generate a number in such a way that under repeated sampling in the long run the proportion of values falling in any range agrees with that prescribed by the population distribution. Of course, we typically don't know the population distribution, as this is what we want to find out about in a statistical investigation. Still, there are many instances where we want to pretend that we do know it and simulate from this distribution; for example, perhaps we want to consider the effect of various choices of population distribution on the sampling distribution of some statistic of interest.

There are computer algorithms that allow us to do this for a variety of distributions. In Minitab, this is accomplished using the Calc ▶ Random Data command. For example, suppose that we want to simulate the tossing of a fair coin (a coin where head and tail are equally likely as outcomes). The Calc ▶ Random Data ▶ Bernoulli command together with the dialog box of Display 3.2.1 generates a sample of 100 from the Bernoulli(.5) distribution and places these values in C1. A random variable has a Bernoulli(p) distribution if the probability the variable equals 1—success—is p and the probability the variable equals 0—failure—is $1 - p$. So to generate a sample of n from the Bernoulli(p) distribution, we put n in the Number of rows to generate box and p in the Event probability box. In such a case, we are simulating the tossing of a coin that produces a head on a single toss with probability p; i.e., the long-run proportion of heads that we observe in repeated tossing is p. Note that we can generate m samples of size n by putting m distinct columns in the Store in column(s) box.

Often, a normal distribution with some particular mean and standard deviation is considered a reasonable assumption for the distribution of a measurement in a population. For example, the Calc ▶ Random Data ▶ Normal command together with the dialog box of Display 3.2.2 generates a sample of 200 from the $N(3.0, 0.4)$ distribution and places this sample in C1. To generate a sample of n from the $N(\mu, \sigma)$ distribution, we put n in the Number of rows to generate box, μ in the Mean box, and σ in the Standard deviation box.



Display 3.2.1: Dialog box for generating a sample form a Bernoulli distribution.



Display 3.2.2: Dialog box for generating a sample of 200 from a $N(3.0, 0.4)$ distribution.

The general syntax of the corresponding session command **random** is

random V into $E_1 \ldots E_m$

and this puts a sample of size V into each of the columns E_1, ..., E_m, according to the distribution specified by the subcommand. For example,

```
MTB > random 100 c1;
SUBC> bernoulli .5.
```

simulates the tossing of a fair coin 100 times and places the results in C1 using the **bernoulli** subcommand. If no subcommand is provided, this distribution is taken to be the $N(0, 1)$ distribution. The command

```
MTB > random 200 c1;
SUBC> normal mu=2.1 sigma=3.3.
```

generates a sample of 200 from the $N(2.1, 3.3)$ distribution using the **normal** subcommand. There are a number of other subcommands specifying distributions, and we refer the reader to **help** for a description of these.

3.3 Exercises

If your version of Minitab places restrictions such that the value of the simulation sample size N requested in these problems is not feasible, then substitute a more appropriate value. Be aware, however, that the accuracy of your results is dependent on how large N is.

1. Enter 10 names into column C1 in alphabetical order and then generate a random permutation of the names storing the result in C2.

2. Use Minitab to generate a random sample of 50 from $\{1, 2, \ldots, 100\}$. Next generate a sample of 50 with replacement. Explain the difference between these samples.

3. Use the following methodology to generate a sample of 20 from a population of 100,000. First, put the values 0–9 in each of C1–C5. Next, use sampling with replacement to generate 50 values from C1, and put the results in C6. Do the same for each of C2–C5 and put the results in C7–C10 (don't generate from these columns simultaneously). Create a single column of numbers using the digits in C6–C10 as the digits in the numbers. Pick out the first unique 20 entries as labels for the sample. If you do not obtain 20 unique values, repeat the process until you do. Why does this work?

4. Suppose you wanted to carry out stratified sampling where there are three strata, with the first stratum containing 500 elements, the second stratum containing 400 elements, and the third stratum containing 100 elements. Generate a stratified sample with 50 elements from the first stratum, 40 elements from the second stratum, and 10 elements from the third stratum. When the strata sample sizes are the same proportion of the total sample size as the strata population sizes are of the total population size this is called *proportional sampling*.

5. Suppose we have an urn containing 100 balls with 20 labeled 1, 50 labeled 2, and 30 labeled 3. Using sampling with replacement, generate a sample of size 1000 from this distribution employing the Calc ▶ Random Data command to generate the sample directly from the relevant population distribution. Use the Stat ▶ Tables ▶ Tally Individual Variables command to record the count of each label in the sample.

6. Suppose we toss a coin n times and then estimate the probability p of getting a head on a single toss by the proportion of heads in the sample $\hat{p}$. Carry out a simulation study with $N = 1000$ of the sampling distribution of $\hat{p}$ for $n = 5, 10, 20$ and for $p = .5, .75, .95$. In particular, calculate the empirical distribution functions and plot the histograms. Comment on your findings.

7. Carry out a simulation study with $N = 2000$ of the sampling distribution of the sample standard deviation when sampling from the $N(0, 1)$ distribution based on a sample of size $n = 5$. In particular, plot the histogram using cutpoints 0, 1.5, 2.0 2.5, 3.0 5.0. Repeat this for the sample coefficient of variation (sample standard deviation divided by the sample mean) using the cutpoints -10, -9, ..., 0, ..., 9, 10. Comment on the shapes of the histograms relative to an $N(0, 1)$ density curve.

Chapter 4

Probability: The Study of Randomness

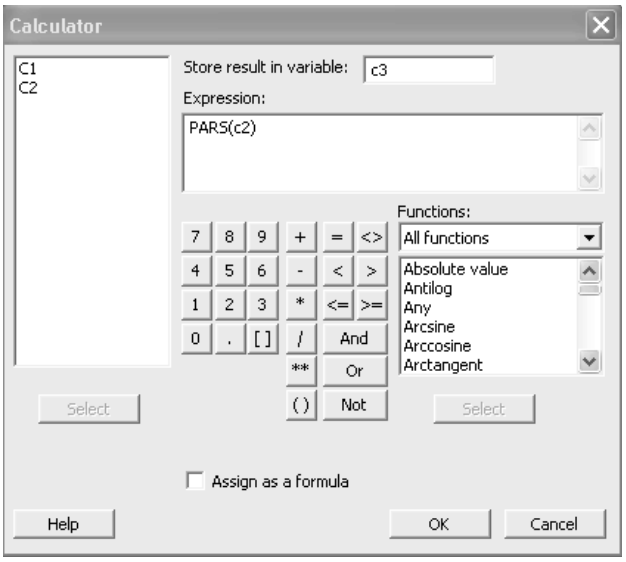
Probability theory underlies the powerful computational methodology known as simulation, which we introduced in Chapter 3. Simulation has many applications in probability and statistics and also in many other fields, such as engineering, chemistry, physics, and economics.

4.1 Basic Probability Calculations

The calculation of probabilities for random variables can often be simplified by tabulating the cumulative distribution function. Also, means and variances are easily calculated using component-wise column operations in Minitab. For example, suppose we have the probability distribution

x	1	2	3	4
probability	.1	.2	.3	.4

in columns C1 and C2, with the values in C1 and the probabilities in C2. The Calc ▶ Calculator command with the dialog box as in Display 4.1.1 computes the cumulative distribution function in C3 using Partial Sums.



Display 4.1.1: Dialog box for computing partial sums of entries in C2 and placing these sums in C3.

Printing C1 and C3 gives

```
Row   C1    C3
1      1   0.1
2      2   0.3
3      3   0.6
4      4   1.0
```

in the Session window. We can also easily compute the mean and variance of this distribution. For example, the session commands

```
MTB > let c4=c1*c2
MTB > let c5=c1*c1*c2
MTB > let k1=sum(c4)
MTB > let k2=sum(c5)-k1*k1
MTB > print k1 k2
K1 3.00000
K2 1.00000
```

calculate the mean and variance and store these in K1 and K2, respectively. The mean is 3 and the variance is 1. Of course, we can also use Calc ▶ Calculator to do these calculations. In presenting more extensive computations, it is somewhat easier to list the appropriate session commands, as we will do subsequently. However, this is not to be interpreted as the required way to do these computations, as it is obvious that the menu commands can be used as well. Use whatever you find most convenient.

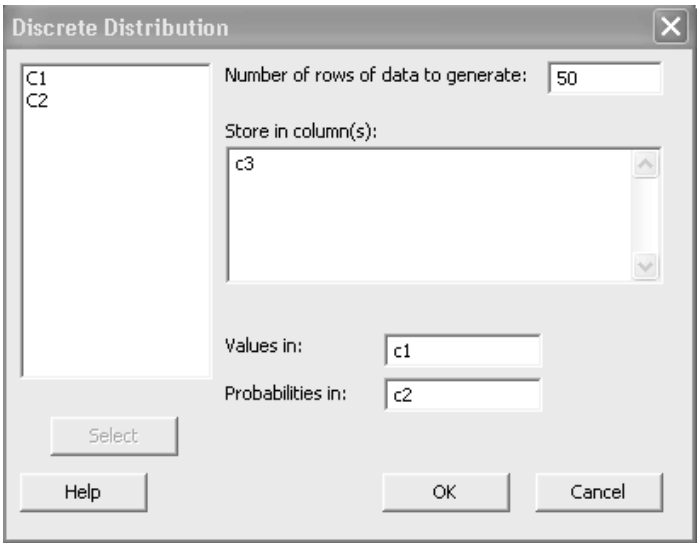
4.2 More on Sampling from Distributions

As we saw in Section 3.2, Minitab includes algorithms for generating from many probability distributions using Calc ▶ Random Data. This menu command produces a drop-down list that includes the normal, binomial, Chi-square, F, t, uniform, and many other distributions. Clicking on one of these names results in a dialog box with entries to be filled in further specifying the distribution and the size of the sample.

For example, we can generate from one particularly important class of probability distributions using Calc ▶ Random Data ▶ Discrete. These probability distributions are concentrated on a finite number of values. To illustrate this, suppose we have the following values in C1 and C2.

Row	C1	C2
1	-1	0.3
2	2	0.2
3	3	0.4
4	10	0.1

Here, C1 contains the possible values of an outcome, and C2 contains the probabilities that each of these values is obtained, so, for example, $P(\{-1\}) = .3, P(\{2\}) = .2$, etc. The dialog box of Display 4.2.1 generates a sample of 50 from this discrete distribution and stores the sample in C3.

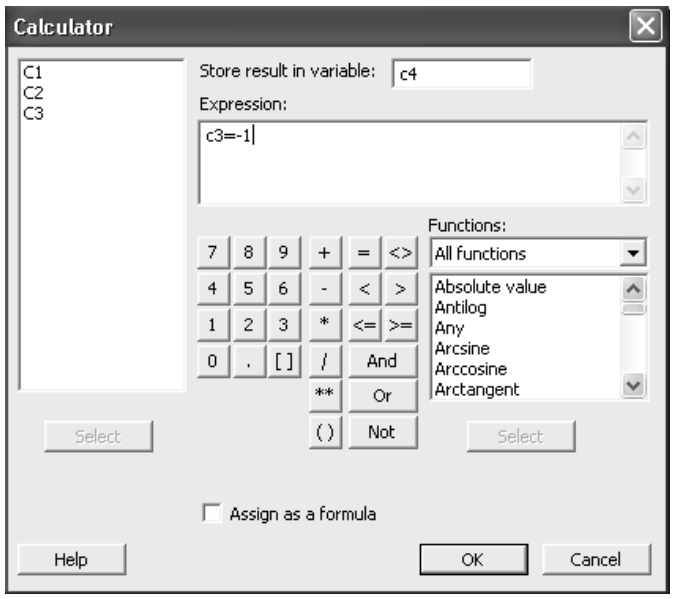


Display 4.2.1: Dialog box for generating a sample from a discrete distribution with values in C1 and probabilities in C2 and storing the sample in C3.

It is an interesting exercise to check that the algorithms Minitab is using are in fact producing samples appropriately. There are a variety of things one could check, but perhaps the simplest is to check that the long-run relative frequencies are correct. So in the example of this section, we want to make sure that, as

we increase the size of the sample, the relative frequencies of $-1, 2, 3, 10$ in the sample are getting closer to .3, .2, .4, and .1, respectively. Note that it is not guaranteed that as we increase the sample size that the relative frequencies get closer monotonically to the corresponding probabilities, but inevitably this must be the case.

First, we generated a sample of size 100 from this distribution and stored the values in C3 as in Display 4.2.1. Next, we recorded a 1 in C4 whenever the corresponding entry in C3 was -1 and recorded a 0 in C4 otherwise. To do this, we used the Calc ▶ Calculator command with dialog box as shown in Display 4.2.2.



Display 4.2.2: Dialog box to record the incidence of a -1 in C3.

It is clear that the mean of C4 is the relative frequency of -1 in the sample. We calculated this mean using Calc ▶ Column Statistics, as discussed in I.10.4, which gave the output

```
Mean of C4 = 0.33000
```

in the Session window. Repeating this with a sample of size 1000, we obtained

```
Mean of C4 = 0.28100
```

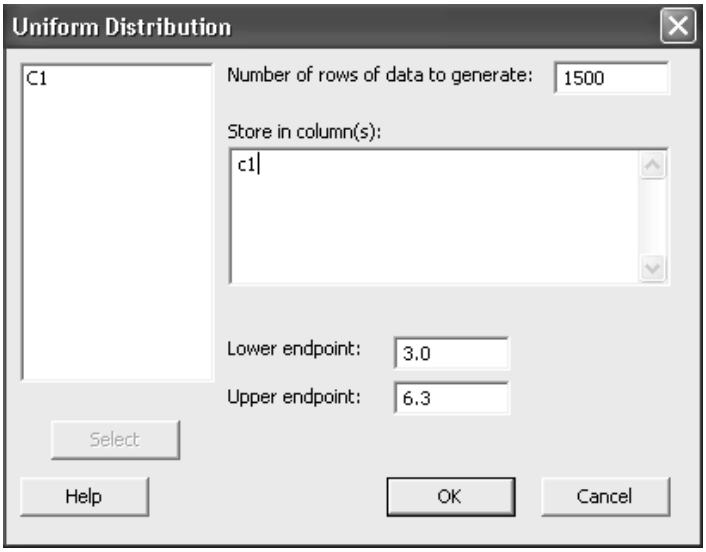
which we can see is a bit closer to the true value of .3. Repeating this with a sample of size $10,000$ from this distribution, we obtained

```
Mean of C4 = 0.29300
```

which is closer still. It would appear that the relative frequency of -1 is indeed converging to .3.

We can generate a randomly chosen point from the line interval (a, b), where $a < b$, using Calc ▶ Random Data ▶ Uniform. For example, the dialog box

of Display 4.2.3 generates a sample of 1500 from the uniform distribution on the interval $(3.0, 6.3)$. With this distribution, the probability of any subinterval (c, d) of (a, b) is given by $(d - c) / (b - a)$, i.e., the length of (c, d) over the length of (a, b). Of course, we can estimate this probability by just counting the number of times the generated response falls in the interval (c, d) and dividing this by the total sample size. For example, using the outcomes from the dialog box of Display 4.2.3 and estimating the probability of the interval $(4, 5)$, we get the relative frequency 0.30867, which is close to the true value of $(5 - 4) / (6.3 - 3) = 0.30303$.



Display 4.2.3: Dialog box for generating a sample of 1500 from a Uniform$(3, 6.3)$ distribution and storing the sample in C3.

We can generalize this to generate from a point randomly chosen from a rectangle $(a, b) \times (c, d)$, i.e., the set of all points (x, y) such that $a < x < b, c < y < d$. If we want a sample of n from this distribution, we generate a sample $x_1, \ldots, x_n$ from the uniform on (a, b) and also generate a sample $y_1, \ldots, y_n$ from the uniform distribution on (c, d). Then $(x_1, y_1), \ldots, (x_n, y_n)$ is a sample of n from the uniform distribution on $(a, b) \times (c, d)$. We can approximate the probability of a random pair (x, y) falling in any subset $A \subset (a, b) \times (c, d)$ by computing the relative frequency of A in the sample.

The **random** command is the session command for carrying out simulations in Minitab. For example, the subcommand

uniform V_1 V_2

specifies the continuous uniform distribution on the interval (V_1, V_2); i.e., subintervals of the same length have the same probability of occurring. If we have placed a discrete probability distribution in column E_2, on the values in column E_1, the subcommand

discrete E_1 E_2

generates a sample from this distribution.

4.3 Simulation for Approximating Probabilities

As previously noted, simulation can be used to approximate probabilities. For a variety of reasons, these simulations are most easily presented using session commands, but it is clear that we can replace each step by the appropriate menu command.

For example, suppose we are asked to calculate

$$P(.1 \leq X_1 + X_2 \leq .3)$$

when X_1, X_2 are both independent and follow the uniform distribution on the interval $(0, 1)$. The session commands

```
MTB > random 1000 c1 c2;
SUBC> uniform 0 1.
MTB > let c3=c1+c2
MTB > let c4 = .1<=c3 and c3<=.3
MTB > let k1=sum(c4)/n(c4)
MTB > print k1
K1 0.0400000
MTB > let k2=sqrt(k1*(1-k1)/n(c4))
MTB > print k2
K2 0.00619677
MTB > let k3=k1-3*k2
MTB > let k4=k1+3*k2
MTB > print k3 k4
K3 0.0214097
K4 0.0585903
```

generate $N = 1000$ independent values of X_1, X_2 and place these values in C1 and C2, respectively, then calculate the sum $X_1 + X_2$ and put these values in C3. Using the comparison operators discussed in I.10.3, a 1 is recorded in C4 every time $.1 \leq X_1 + X_2 \leq .3$ is true and a 0 is recorded there otherwise. We then calculate the proportion of 1's in the sample as K1, and this is our estimate $\hat{p}$ of the probability. We will see later that a good measure of the accuracy of this estimate is the *standard error of the estimate*, which in this case is given by

$$\sqrt{\hat{p}\,(1-\hat{p})\,/N},$$

and this is computed in K2. Actually, we can feel fairly confident that the true value of the probability is in the interval

$$\hat{p} \pm 3\sqrt{\hat{p}\,(1-\hat{p})\,/N},$$

which, in this case, equals the interval $(0.0214097, 0.0585903)$. So we know the true value of the probability with reasonable accuracy. As the simulation size N increases, the Law of Large Numbers says that $\hat{p}$ converges to the true value of the probability.

4.4 Simulation for Approximating Means

The means of distributions can be approximated using simulations in Minitab. For example, suppose X_1, X_2 are both independent and follow the uniform distribution on the interval $(0, 1)$ and that we want to calculate the mean of $Y = 1/(1 + X_1 + X_2)$. We can approximate this in a simulation. The session commands

```
MTB > random 1000 c1 c2;
SUBC> uniform 0 1.
MTB > let c3=1/(1+c1+c2)
MTB > let k1=mean(c3)
MTB > let k2=stdev(c3)/sqrt(n(c3))
MTB > print k1 k2
K1 0.521532
K2 0.00375769
MTB > let k3=k1-3*k2
MTB > let k4=k1+3*k2
MTB > print k3 k4
K3 0.510259
K4 0.532805
```

generate $N = 1000$ independent values of X_1, X_2 and place these values in C1, C2, then calculate $Y = 1/(1 + X_1 + X_2)$ and put these values in C3. The mean of C3 is stored in K1, and this is our estimate of the mean value of Y. As a measure of how accurate this estimate is, we compute the standard error of the estimate, which is given by the standard deviation divided by the square root of the simulation sample size N. Again, we can feel fairly confident that the interval given by the estimate plus or minus 3 times the standard error of the estimate contains the true value of the mean. In this case, this interval is given by $(0.510259, 0.532805)$, and so we know this mean with reasonable accuracy. As the simulation size N increases, the Law of Large Numbers says that the approximation converges to the true value of the mean.

4.5 Exercises

If your version of Minitab places restrictions such that the value of the simulation sample size N requested in these problems is not feasible, then substitute a more appropriate value. Be aware, however, that the accuracy of your results is dependent on how large N is.

1. Suppose we have the probability distribution

x	1	2	3	4	5
probability	.15	.05	.33	.37	.10

on the values 1, 2, 3, 4, and 5. Calculate the mean and variance of this distribution. Suppose that three independent outcomes (X_1, X_2, X_3) are generated from this distribution. Compute the probability that $1 < X_1 \leq 4, 2 \leq X_2$ and $3 < X_3 \leq 5$.

2. Suppose we have the probability distribution

x	1	2	3	4	5
probability	.15	.05	.33	.37	.10

on the values 1, 2, 3, 4, and 5. Using Minitab, verify that this is a probability distribution. Make a bar chart (probability histogram) of this distribution. Generate a sample of size 1000 from this distribution and plot a relative frequency histogram for the sample.

3. Indicate how you would simulate the game of roulette using Minitab. Based on a simulation of $N = 1000$, estimate the probability of getting red and a multiple of 3.

4. A probability distribution is placed on the integers 1, 2, ..., 100, where the probability of integer i is c/i^2. Determine c so that this is a probability distribution. What is the 90th percentile? Generate a sample of 20 from the distribution.

5. Suppose an outcome is random on the square $(0, 1) \times (0, 1)$. Using simulation, approximate the probability that the first coordinate plus the second coordinate is less than .75 but greater than .25.

6. Generate a sample of 1000 from the uniform distribution on the unit disk $D = \left\{(x, y) : x^2 + y^2 \leq 1\right\}$.

7. The expression e^{-x} for $x > 0$ is the density curve for what is called the Exponential(1) distribution. Plot this density curve in the interval from 0 to 10 using an increment of .1. The Calc ▶ Random Data ▶ Exponential command can be used to generate from this distribution by specifying the Mean as 1 in the ensuing dialog box. Generate a sample of 1000 from this distribution and estimate its mean. Approximate the probability that a value generated from this distribution is in the interval (1,2). The general Exponential(λ) has a density curve given by $\lambda^{-1}e^{-x/\lambda}$ for $x > 0$ and where $\lambda > 0$ is the mean. Repeat the simulation with mean $\lambda = 3$. Comment on the values of the estimated means.

8. Suppose you carry out a simulation to approximate the mean of a random variable X and you report the value 1.23 with a standard error of .025.

If you are asked to approximate the mean of $Y = 3 + 5X$, do you have to carry out another simulation? If not, what is your approximation, and what is the standard error of this approximation?

9. Suppose that a random variable X follows an $N(3, 2.3)$ distribution. Subsequently, conditions change and no values smaller than -1 or bigger than 9.5 can occur; i.e., the distribution is conditioned to the interval $(-1, 9.5)$. Generate a sample of 1000 from the truncated distribution, and use the sample to approximate its mean.

10. Suppose that X is a random variable and follows an $N(0, 1)$ distribution. Simulate $N = 1000$ values from the distribution of $Y = X^2$, and plot these values in a histogram with cutpoints 0, .5, 1, 1.5, ..., 15. Approximate the mean of this distribution. Generate Y directly from its distribution, which is known to be a Chi-square(1) distribution. In general, the Chi-square(k) distribution can be generated from via the command Calc ▶ Random Data ▶ Chi-Square, where k is specified as the Degrees of freedom in the dialog box. Plot the Y values in a histogram using the same cutpoints. Comment on the two histograms. Note that you can plot the density curve of these distributions using Calc ▶ Probability Distributions ▶ Chi-Square and evaluating the probability density at a range of points as we discussed in II.2 for the normal distribution.

11. If X_1 and X_2 are independent random variables with X_1 following a Chi-square(k_1) distribution and X_2 following a Chi-square(k_2) distribution, then it is known that $Y = X_1 + X_2$ follows a Chi-square($k_1 + k_2$) distribution. For $k_1 = 1$, $k_2 = 1$, verify this empirically by plotting a density histogram with cutpoints 0, .5, 1, 1.5, ..., 15, based on two samples of size $N = 1000$ from the Chi-square(1) distribution and compare this with a plot of the Chi-square(2) density curve.

12. If X_1 and X_2 are independent random variables with X_1 following an $N(0, 1)$ distribution and X_2 following a Chi-square(k) distribution, then it is known that

$$Y = \frac{X_1}{\sqrt{X_2/k}}$$

follows a Student(k) distribution. The Student(k) distribution can be generated from using the command Calc ▶ Random Data ▶ t, where k is the Degrees of freedom and must be specified in the dialog box. For $k = 3$, verify this result empirically by plotting histograms with cutpoints -10, -9, ..., 9, 10, based on simulations of size $N = 1000$; i.e., generate 1000 values of (X_1, X_2), plot a density histogram of the Y values, and compare this with a plot of the density curve of a Student(3) distribution.

13. If X_1 and X_2 are independent random variables with X_1 following a Chi-square(k_1) distribution and X_2 following a Chi-square(k_2) distribution, then it is known that

$$Y = \frac{X_1/k_1}{X_2/k_2}$$

follows an $F(k_1, k_2)$ distribution. The $F(k_1, k_2)$ distribution can be generated from using the subcommand Calc ▶ Random Data ▶ F, where k_1 is the Numerator degrees of freedom and k_2 is the Denominator degrees of freedom, both of which must be specified in the dialog box. For $k_1 = 1$, $k_2 = 1$, verify this empirically by plotting histograms with cutpoints 0, .5, 1, 1.5, ..., 15, based on simulations of size $N = 1000$; i.e., generate 1000 values of (X_1, X_2), plot a density histogram of the Y values, and compare this with a plot of the density curve of a $F(1, 1)$ distribution.

Chapter 5

Sampling Distributions

New Minitab command discussed in this chapter

Calc ▶ Probability Distributions ▶ Binomial

Once data have been collected, they are analyzed using a variety of statistical techniques. Virtually all of these involve computing *statistics* that measure some aspect of the data concerning questions we wish to answer. The answers determined by these statistics are subject to the uncertainty caused by the fact that we typically do not have the full population but only a sample from the population. As such, we have to be concerned with the variability in the answers when different samples are obtained. This leads to a concern with the *sampling distribution* of a statistic.

Sometimes, the sampling distribution of a statistic can be worked out exactly through various mathematical techniques; for example, it can be shown that the number of 1's in a sample of n from a Bernoulli(p) distribution follows a Binomial(n, p) distribution. Often, however, this is not possible, and we must resort to approximations. One approximation technique is to use simulation. Sometimes, however, the statistics we are concerned with are averages, and, in such cases, the central limit theorem justifies approximating the sampling distribution via an appropriate normal distribution.

5.1 The Binomial Distribution

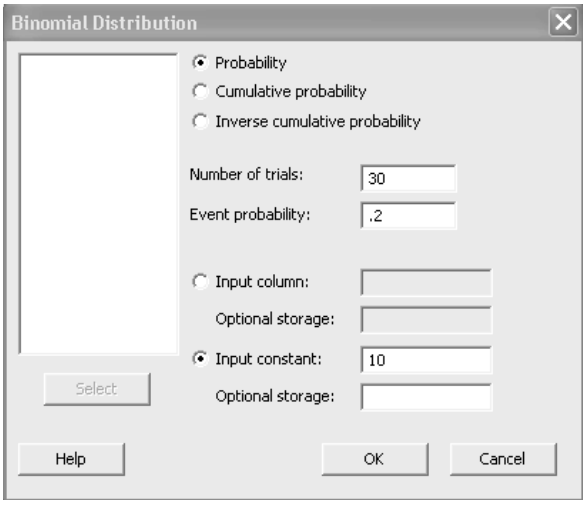
Suppose that $X_1, \ldots, X_n$ is a sample from the Bernoulli(p) distribution; i.e., $X_1, \ldots, X_n$ are independent realizations, where each X_i takes the value 1 or 0 with probabilities p and $1 - p$, respectively. The random variable $Y = X_1 + \cdots + X_n$ equals the number of 1's in the sample and follows a Binomial(n, p) distribution. Therefore, Y can take on any of the values $0, 1, \ldots, n$ with positive probability. In fact, an exact formula can be derived for these probabilities;

namely, $P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$ is the probability that Y takes the value k for $0 \le k \le n$. When n and k are small, this formula could be used to evaluate this probability, but it is almost always better to use software like Minitab to do it, and when these values are not small, it is necessary. Also, we can use Minitab to compute the Binomial(n, p) cumulative probability distribution—the probability contents of intervals $(-\infty, x]$ and the inverse cumulative distribution—quantiles of the distribution.

For individual probabilities, we use the Calc ▶ Probability Distributions ▶ Binomial command. For example, suppose we have a Binomial $(30, .2)$ distribution and want to compute the probability $P(Y = 10)$. This command, with the dialog box as in Display 5.1.1, produces the output (note Minitab uses the notation p instead of p for the probability of success)

```
Binomial with n = 30 and p = 0.200000
   x          P( X = x )
10.00       0.0354709
```

in the Session window, i.e., $P(Y = 10) = 0.0354709$.

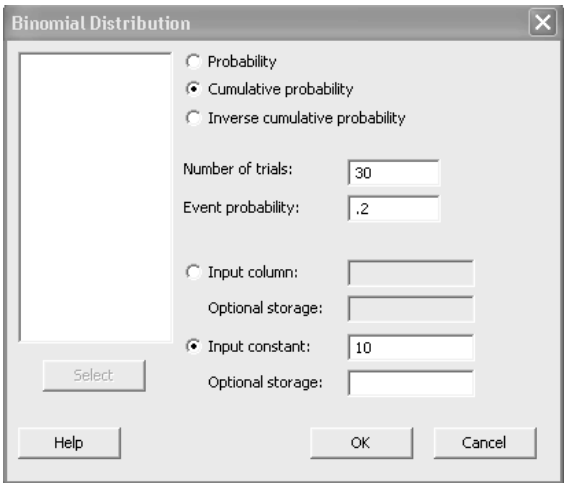


Display 5.1.1: Dialog box for Binomial(n, p) probability calculations.

If we want to compute the probability of getting 10 or fewer successes (this is the probability of the interval $(-\infty, 10]$) we can use the Calc ▶ Probability Distributions ▶ Binomial command with the dialog box as in Display 5.1.2. This produces the output

```
Binomial with n = 30 and p = 0.200000
  x         P( X <= x )
10.00       0.974384
```

in the Session window, i.e., $P(Y \le 10) = 0.974384$.

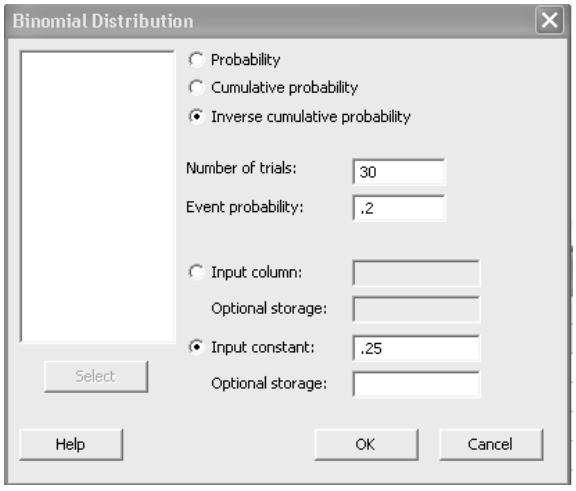


Display 5.1.2: Dialog box for computing cumulative probabilities for the Binomial(n, p) distribution.

Suppose we want to compute the first quartile of this distribution. The Calc ▶ Probability Distributions ▶ Binomial command, with the dialog box as in Display 5.1.3, produces the output

```
Binomial with n = 30 and p = 0.200000
 x    P( X <= x )    x      P( X <= x )
 3      0.122711     4        0.255233
```

in the Session window. This gives the values x that have cumulative probabilities just smaller and just larger than the value requested. Recall that with a discrete distribution, such as the Binomial(n, p), we will not in general be able to obtain an exact quantile.



Display 5.1.3 Dialog box for computing percentiles of the Binomial(n, p) distribution.

These commands can operate on all the values in a column simultaneously. This is very convenient if you should want to tabulate or graph the probability function, cumulative distribution function, or inverse distribution function.

The corresponding session commands are **pdf** (for calculating the probability function), **cdf** (for calculating the cdf), and **invcdf** (for calculating the inverse cdf) used with the **binomial** subcommand. For example,

```
MTB > pdf 10;
SUBC> binomial 30 .2.
```

outputs $P(Y = 10)$ when Y has the Binomial$(30, .2)$ distribution.

5.2 Simulating Sampling Distributions

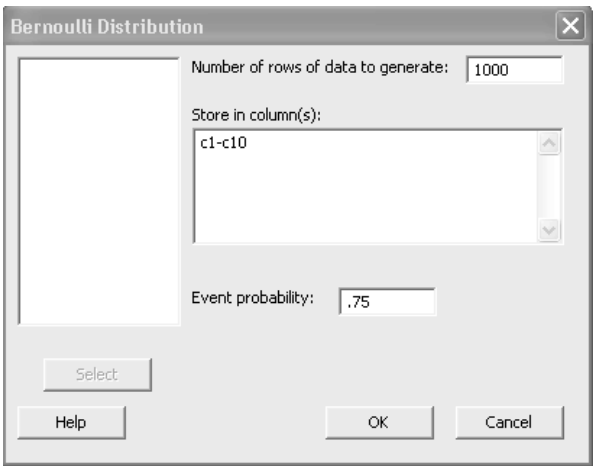
First, we consider an example where we know the exact sampling distribution. Suppose we flip a possibly biased coin n times and want to estimate the unknown probability p of getting a head. The natural estimate is $\hat{p}$ the proportion of heads in the sample. We would like to assess the sampling behavior of this statistic in a simulation. To do this, we choose a value for p, then generate N samples from the Bernoulli distribution of size n; for each of these compute $\hat{p}$, look at the empirical distribution of these N values, perhaps plotting a histogram as well. The larger N is the closer the empirical distribution and histogram will be to the true sampling distribution of $\hat{p}$.

Note that there are two sample sizes here: the sample size n of the original sample the statistic is based on, which is fixed, and the *simulation* sample size N, which we can control. This is characteristic of all simulations. Sometimes, using more advanced analytical techniques we can determine N so that the sampling distribution of the statistic is estimated with some prescribed accuracy. These methods are referred to as setting the sample size. Another method is to increase N until we see the results stabilize. This is sometimes the only way available, but caution should be shown as it is easy for simulation results to be very misleading if the final N is too small.

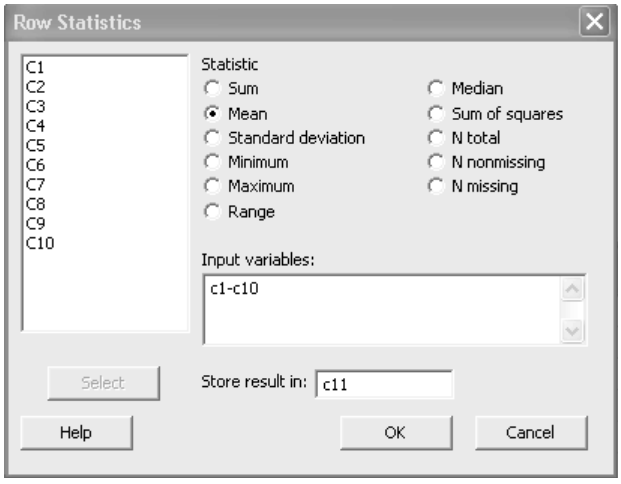
We illustrate a simulation to determine the sampling distribution of $\hat{p}$ when sampling from a Bernoulli(.75) distribution. For this, we use the commands Calc ▶ Random Data ▶ Bernoulli, Calc ▶ Row Statistics, and Stat ▶ Tables ▶ Tally Individual Variables, with the dialog boxes given by Displays 5.2.1, 5.2.2, and 5.2.3, respectively, to produce the output

```
Summary Statistics for Discrete Variables
 C11 CumPct
 0.3   0.40
 0.4   2.20
 0.5   7.60
 0.6  23.10
 0.7  47.70
 0.8  78.00
 0.9  94.70
 1.0 100.00
```

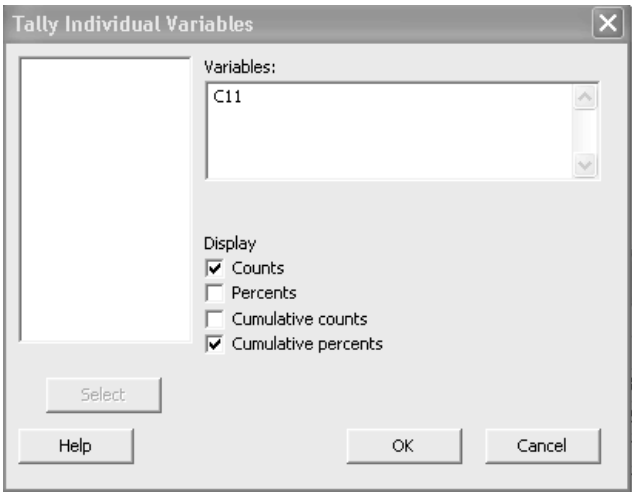
in the Session window. Here we have generated $N = 1000$ samples of size $n = 10$ from the Bernoulli(.75) distribution; i.e., we simulated the tossing of this coin 10,000 times, and we placed the results in the rows of columns C1–C10 using Calc ▶ Random Data ▶ Bernoulli. The proportion of heads $\hat{p}$ in each sample is computed and placed in C11 using Calc ▶ Row Statistics. Note that a mean of values equal to 0 or 1 is just the proportion of 1's in the sample. Finally, we used Stat ▶ Tables ▶ Tally Individual Variables to compute the empirical distribution function of these 1000 values of $\hat{p}$. For example, this says 78% of these values were .8 or smaller and there were no instances smaller than .3. In Display 5.2.4, we have plotted a density histogram of the 1000 values of $\hat{p}$, and this gives a rough idea of the shape of the sampling distribution.



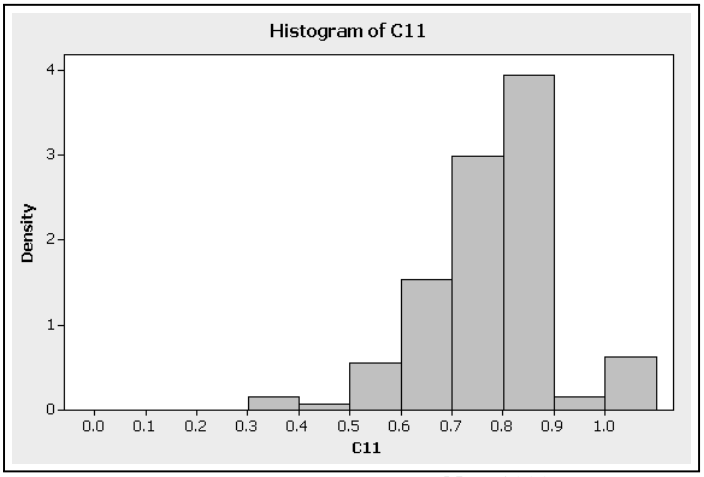
Display 5.2.1: Dialog box for generating 10 columns of 1000 Bernoulli(.75) values.



Display 5.2.2: Dialog box for computing the proportion of 1's in each of the 1000 samples of size 10.



Display 5.2.3: Dialog box for computing the empirical distribution function of $\hat{p}$.



Display 5.2.4: Density histogram of simulation of $N = 1000$ values of $\hat{p}$ based on a sample of size $n = 10$ from the Bernoulli(.75) distribution.

The corresponding session commands for this simulation are

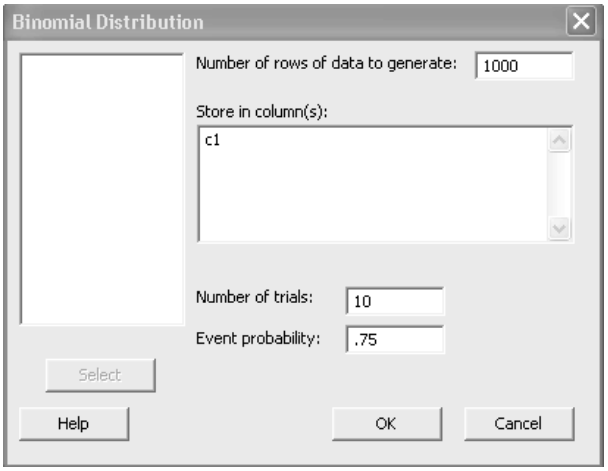
```
MTB > random 1000 c1-c10;
SUBC> bernoulli .75.
MTB > rmean c1-c10 c11
MTB > tally c11;
SUBC> cumpcts.
```

and these might seem like an easier way to implement the simulation.

As mentioned previously, the sampling distribution of $\hat{p}$ can be determined exactly; i.e., there are formulas to determine this, so really there is no need for a simulation in this case. Still, it illustrates how such a simulation proceeds in more general circumstances.

Furthermore, we can simulate directly from the sampling distribution of $\hat{p}$, so this simulation can be made much more efficient. In effect, this entails using the Calc ▶ Random Data ▶ Binomial command with dialog box as in Display 5.2.5 and dividing each entry in C1 by 10. This generates $N = 1000$ values of $\hat{p}$ but uses a much smaller number of cells. Still, there are many statistics for which this kind of efficiency reduction is not available, and, to get some idea of what their sampling distribution is like, we must resort to the more brute force form of simulation of generating directly from the population distribution.

Sometimes, more sophisticated simulation techniques are needed to get an accurate assessment of a sampling distribution. Within Minitab, there are programming techniques, which we do not discuss in this manual, that can be applied in such cases. For example, it is clear that if our simulation required the generation of 10^6 cells (and this is not at all uncommon for some harder problems), the simulation approach we have described would not work, as the worksheet would be too large.



Display 5.2.5: Dialog box for generating 1000 values from the sampling distribution of $10\hat{p}$ using the Binomial$(10, .75)$ distribution.

5.3 Exercises

If your version of Minitab places restrictions such that the value of the simulation sample size N requested in these problems is not feasible, then substitute a more appropriate value. Be aware, however, that the accuracy of your results is dependent on how large N is.

1. Calculate all the probabilities for the Binomial$(5, .4)$ distribution and the Binomial$(5, .6)$ distribution. What relationship do you observe? Can you explain this and state a general rule?

2. Compute all the probabilities for a Binomial$(5, .8)$ distribution and use

these to directly calculate the mean and variance. Verify your answers using the formulas provided in your text.

3. Compute and plot the probability and cumulative distribution functions of the Binomial$(10, .2)$ and the Binomial$(10, .5)$ distributions. Comment on the shapes of these distributions.

4. Generate 1000 samples of size 10 from the Bernoulli(.3) distribution. Compute the proportion of 1's in each sample and compute the proportion of samples having no 1's, one 1, two 1's, etc. Compute what these proportions would be in the long run and compare.

5. Carry out a simulation study with $N = 1000$ of the sampling distribution of $\hat{p}$ for $n = 5, 10, 20$ and for $p = .5, .75, .95$. In particular, calculate the empirical distribution functions and plot the histograms. Comment on your findings.

6. Suppose that $X_1, X_2, \ldots$ are independent realizations from the Bernoulli(p) distribution; i.e., each X_i takes the value 1 or 0 with probabilities p and $1 - p$, respectively. If the random variable Y counts the number of tosses until we obtain the first head in a sequence of independent tosses $X_1, X_2, X_3, \ldots$, then Y has a Geometric(p) distribution. The probability function for this distribution is given by

$$P(Y = y) = (1 - p)^{y-1} p$$

for $y = 1, 2, \ldots$. Plot the probability function for the Geometric(.5) distribution for the values $y = 1, \ldots, 10$. Do the same for the Geometric(.1) distribution. What do you notice?

7. Using methods for summing geometric sums, the cumulative distribution function of the Geometric(p) distribution (see Exercise II.5.6) is given by $P(Y \leq y) = 1-(1-p)^y$. Plot the cumulative distribution function for the Geometric(.5) and Geometric(.1) distribution for the values $y = 1, \ldots, 10$. What do you notice?

8. To randomly generate from the Geometric(p) distribution (see Exercise II.5.6), we can repeatedly generate from a Bernoulli(p) and count how many times we did this until the first 1 appeared. Using Minitab generate a sample of 1000 from the Geometric(.5) distribution. Plot the sample in a proportion histogram.

9. Carry out a simulation study, with $N = 2000$, of the sampling distribution of the sample standard deviation when sampling from the $N(0, 1)$ distribution, based on a sample of size $n = 5$. In particular, plot the histogram using cutpoints 0, 1.5, 2.0 2.5, 3.0 5.0. Repeat this for the sample coefficient of variation (sample standard deviation divided by the sample mean) using the cutpoints -10, -9, ..., 0, ..., 9, 10. Comment on the shapes of the histograms relative to a $N(0, 1)$ density curve.

10. Generate $N = 1000$ samples of size $n = 5$ from the $N(0, 1)$ distribution. Record a histogram for $\bar{x}$ using the cutpoints $-3, -2.5, -2, ..., 2.5, 3.0$. Generate a sample of size $N = 1000$ from the $N(0, 1/\sqrt{5})$ distribution. Plot the histogram using the same cutpoints and compare the histograms. What will happen to these histograms as we increase N?

11. Generate $N = 1000$ values of X_1, X_2, where X_1 follows a $N(3, 2)$ distribution and X_2 follows a $N(-1, 3)$ distribution. Compute $Y = X_1 - 2X_2$ for each of these pairs and plot a histogram for Y using the cutpoints $-20, -15, ..., 25, 30$. Generate a sample of $N = 1000$ from the appropriate distribution of Y and plot a histogram using the same cutpoints.

12. Plot the density curve for the Exponential(3) distribution (see Exercise II.4.7) between 0 and 15 with an increment of .1. Generate $N = 1000$ samples of size $n = 2$ from the Exponential(3) distribution and record the sample means. Standardize the sample of $\bar{x}$ using $\mu = 3$ and $\sigma = 3$. Plot a histogram of the standardized values using the cutpoints -5, -4, ..., 4, 5. Repeat this for $n = 5, 10$. Comment on the shapes of these histograms.

13. Plot the density of the uniform distribution on (0,1). Generate $N = 1000$ samples of size $n = 2$ from this distribution. Standardize the sample of $\bar{x}$ using $\mu = .5$ and $\sigma = \sqrt{1/12}$. Plot a histogram of the standardized values using the cutpoints $-5, -4, ..., 4, 5$. Repeat this for $n = 5, 10$. Comment on the shapes of these histograms.

14. The Weibull(β) has density curve given by $\beta x^{\beta-1} e^{-x^\beta}$ for $x > 0$, where $\beta > 0$ is a fixed constant. Plot the Weibull(2) density in the range 0 to 10 with an increment of .1 using the Calc ▶ Probability _Distributions ▶ Weibull, command. Generate a sample of $N = 1000$ from this distribution using the subcommand Calc ▶ Random Data ▶ Weibull where β is the Shape parameter and the Scale parameter is 1. Plot a probability histogram and compare with the density curve.

Chapter 6

Introduction to Inference

New Minitab commands discussed in this chapter

Stat ▶ Basic Statistics ▶ 1-Sample Z
Power and Sample Size ▶ 1-Sample Z

In this chapter, the basic tools of statistical inference are discussed. There are a number of Minitab commands that aid in the computation of confidence intervals and for carrying out tests of significance.

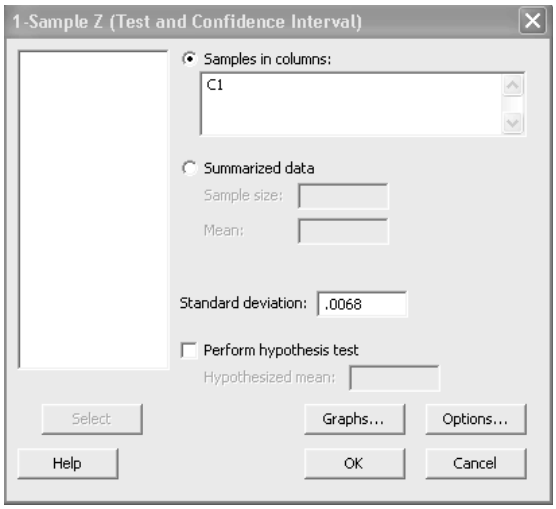
6.1 z Confidence Intervals

The command Stat ▶ Basic Statistics ▶ 1-Sample Z computes confidence intervals of the form $\bar{x} \pm z_{(1+\gamma)/2}\sigma_0/\sqrt{n}$, where γ is prescribed (often $\gamma = 0.95$), σ_0 is known, $\bar{x}$ and n are obtained from the data, and z_α is the α-th percentile of the $N(0,1)$ distribution.

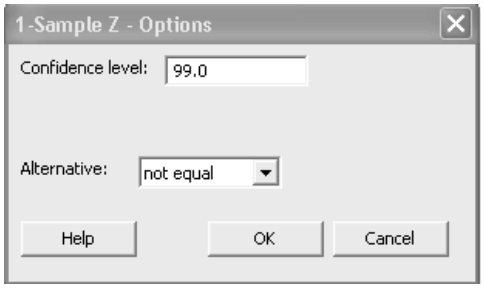
Consider the sample given by $(0.8403, 0.8363, 0.8447)$, which are stored in C1, and suppose that it makes sense to take $\sigma_0 = .0068$. The command Stat ▶ Basic Statistics ▶ 1-Sample Z with the dialog boxes as in Displays 6.1.1 and 6.1.2 produces the output

```
Variable   N    Mean      StDev        SE Mean
  C1       3   0.84043    0.00420       0.00393
      99.% CI
(0.83032, 0.85055)
```

in the Session window. This specifies $(0.83032, 0.85055)$ as a 99% confidence interval for μ. Note that in the dialog box of Display 6.1.1, we specify where the data resides in the Samples in Columns box, the value of σ_0 in the Standard deviation box, and clicked on the Options button to bring up the dialog box in Display 6.1.2. In this dialog box we have specified the 99% confidence level in the Confidence level box.



Display 6.1.1: First dialog box for producing the z confidence interval for μ.



Display 6.1.2: Second dialog box for producing the z confidence interval. Here we specify the confidence level.

The corresponding session command **zinterval** is

zinterval V_1 sigma = V_2 $E_1 \ldots E_m$

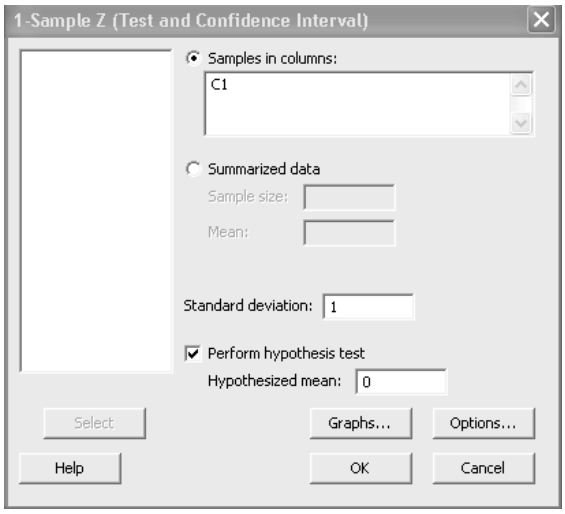
where V_1 is the confidence level and is any value between 1 and 99.99, V_2 is the assumed value of σ, and E_1, ..., E_m are columns of data. A V_1% confidence interval is produced for each column specified. If no value is specified for V_1, the default value is 95%.

6.2 z Tests

The Stat ▶ Basic Statistics ▶ 1-Sample Z command is used when we want to assess hypotheses about the unknown mean μ. Suppose the sample $(2.0, 0.4, 0.7, 2.0, -0.4, 2.2, -1.3, 1.2, 1.1, 2.3)$ is stored in C1, and we are asked to assess the null hypothesis $H_0 : \mu = 0$ and we know that $\sigma_0 = 1$. The Stat ▶ Basic Statistics ▶ 1-Sample Z command—together with the dialog box of Display 6.2.1, where we specified where the data is located, the value of σ_0, and that we want to test $H_0 : \mu = 0$ by placing 0 in the Test mean box—produces the following output:

```
Variable  N     Mean    StDev   SE Mean       99% CI
C1        10    1.020   1.196    0.316    (0.205, 1.835)
  Z     P
3.23  0.001
```

This gives the value of $z = 3.23$ for the z statistic and the P-value equal to 0.001. This is strong evidence against $H_0 : \mu = 0$.



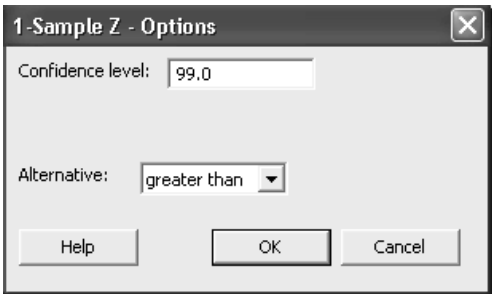
Display 6.2.1: Dialog box for assessing the hypothesis $H_0 : \mu = 0$ using a z test.

Sometimes it is preferred to assess a one-sided hypothesis such as $H_0 : \mu \leq \mu_0$. In this case, the relevant P-value is $P(Z > (\bar{x} - \mu_0)/(\sigma_0/\sqrt{n})) = 1 - \Phi\left((\bar{x} - \mu_0)/(\sigma_0/\sqrt{n})\right)$. Minitab also has the facility for assessing hypotheses such as $H_0 : \mu \leq \mu_0$ or $H_0 : \mu \geq \mu_0$.

Suppose, for the above sample, we are asked to assess the null hypothesis $H_0 : \mu \leq 0$ and we know $\sigma = 1$. The Stat ▶ Basic Statistics ▶ 1-Sample Z command, together with the dialog boxes of Displays 6.2.1 and 6.2.2 (the greater than refers to the values for which the null hypothesis is false), produces the output

```
Variable  N    Mean    StDev   SE Mean
C1       10  1.020    1.196   0.316
99.0% Lower Bound       Z         P
    0.284              3.23     0.001
```

in the Session window. This specifies the P-value for this test as .001, so we have evidence against the null hypothesis. We obtained the dialog box in Display 6.2.2 by clicking on the Options button Display 6.2.1. Here we specified that we want to test the null hypothesis $H_0 : \mu \leq 0$ by selecting "greater than" in the Alternative box. The other choices are "not equal," which selects the null hypothesis $H_0 : \mu = 0$ (the default), and "less than," which selects the null hypothesis $H_0 : \mu \geq 0$.



Display 6.2.2: Dialog box for specifying the kind of test when using a z test.

Note that the P-values for assessing $H_0 : \mu = 0$ and $H_0 : \mu \leq 0$ are both given as 0.001 in the Minitab output, but these have been rounded from the actual values 0.000619 and 0.001238, respectively. In fact, the P-value for the one-sided test is always bigger than the P-value for the two-sided test.

The general syntax of the corresponding session command **ztest** is

ztest V_1 sigma = V_2 $\mathrm{E}_1 \ldots \mathrm{E}_m$

where V_1 is the hypothesized value to be tested, V_2 is the assumed value of σ, and E_1, ..., E_m are columns of data. If no value is specified for V_1, the default is 0. A P-value for the hypothesis is computed for each column. If no **alternative** subcommand is specified, the P-value for $H_0 : \mu = \mathrm{V}_1$ is computed. If the subcommand

```
SUBC> alternative 1.
```

is used, the P-value for $H_0 : \mu \leq \mathrm{V}_1$ is computed. If the subcommand

```
SUBC> alternative -1.
```

is used, the P-value for $H_0 : \mu \geq \mathrm{V}_1$ is computed.

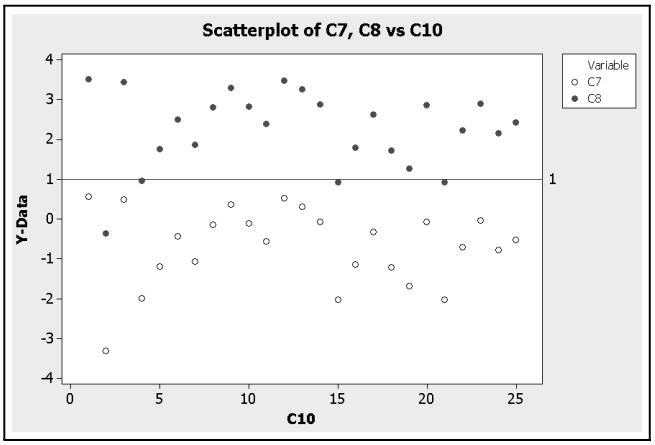
6.3 Simulations for Confidence Intervals

When we are sampling from a $N(\mu, \sigma)$ distribution and know the value of σ, the confidence intervals constructed in Section 6.1 are exact; i.e., in repeated sampling, the long-run proportion of the 95% confidence intervals constructed for an unknown mean μ that will contain the true value of this quantity, is equal to 95%. Of course, any given confidence interval may or may not contain the true value of μ, and, in any finite number of such intervals so constructed, some proportion other than 95% will contain the true value of μ. As the number of intervals increases, however, the proportion covering will go to 95%.

We illustrate this via a simulation study based on computing 90% confidence intervals. The session commands

```
MTB > random 100 c1-c5;
SUBC> normal 1 2.
MTB > rmean c1-c5 c6
MTB > invcdf .95;
SUBC> normal 0 1.
Normal with mean = 0 and standard deviation = 1.00000
 P( X <= x) x
 0.9500 1.6449
MTB > let k1=1.6449*2/sqrt(5)
MTB > let c7=c6-k1
MTB > let c8=c6+k1
MTB > let c9=c7<1 and c8>1
MTB > mean c9
 Mean of C9 = 0.91000
MTB > set c10
DATA> 1:25
DATA> end
MTB > delete 26:100 c7 c8
```

generate 100 random samples of size 5 from the $N(1, 2)$ distribution, place the means in C6, the lower end-point of a 90% confidence interval in C7, and the upper end-point in C8, and record whether or not a confidence interval covers the true value $\mu = 1$ by placing a 1 or 0 in C9, respectively. The mean of C9 is the proportion of intervals that cover, and this is 91%, which is 1% too high. Finally, we plotted the first 25 of these intervals in a plot shown in Display 6.3.1 (note we use the features available in Minitab for producing multiple scatterplots on the same plot to produce this plot). Drawing a solid horizontal line at 1 on the y-axis indicates that most of these intervals do indeed cover the true value $\mu = 1$ (the 2nd, 4th, 15th, and 21st intervals do not contain 1).



Display 6.3.1: Plot of 90% confidence intervals for the mean when sampling from the $N(1, 2)$ distribution with $n = 5$. The lower end-point is denoted by ∘ and the upper end-point is denoted by •.

The simulation just carried out simply verifies a theoretical fact. On the other hand, when we are computing approximate confidence intervals—i.e., we are not sampling necessarily from a normal distribution—it is good to do some simulations from various distributions to see how much reliance we can place in the approximation at a given sample size. The true *coverage probability* of the interval, i.e., the long-run proportion of times that the interval covers the true mean, will not in general be equal to the nominal confidence level. Small deviations are not serious, but large ones are.

6.4 Power Calculations

It is also useful to know in a given context how sensitive a particular test of significance is. By this, we mean how likely it is that the test will lead us to reject the null hypothesis when the null hypothesis is false. This is measured by the concept of the *power* of a test. Typically, a level α is chosen for the P-value at which we would definitely reject the null hypothesis if the P-value is smaller than α. For example, $\alpha = .05$ is a common choice for this level. Suppose that we have chosen the level of .05 for the two-sided z test and we want to evaluate the power of the test when the true value of the mean is $\mu = \mu_1$, i.e., evaluate the probability of getting a P-value smaller than .05 when the mean is μ_1. The two-sided z test with level α rejects $H_0 : \mu = \mu_0$ whenever $2(1-\Phi\left(|(\bar{x}-\mu_0)/(\sigma/\sqrt{n})|\right)) \leq \alpha$ or, equivalently, whenever $|(\bar{x}-\mu_0)/(\sigma/\sqrt{n})| \geq \Phi^{-1}\left(1-\alpha/2\right) = z_{1-\alpha/2}$. For example, if $\alpha = .05$, then $1-\alpha/2 = .975$ and the quantile $z_{.975}$ can be obtained using the command Calc ▶ Probability Distributions ▶ Normal and the inverse distribution function, which gives the output

```
Normal with mean = 0 and standard deviation = 1.00000
 P( X <= x)         x
    0.975       1.95996
```

in the Session window; i.e., the .975 percentile of the $N(0,1)$ distribution is 1.95996.

If $\mu = \mu_1$, then $(\bar{x}-\mu_0)/(\sigma/\sqrt{n})$ is a realized value from the distribution of $Y = (\bar{X}-\mu_0)/(\sigma/\sqrt{n})$ when $\bar{X} \sim N(\mu_1, \sigma/\sqrt{n})$. Therefore, Y follows a $N((\mu_1-\mu_0)/(\sigma/\sqrt{n}), 1)$ distribution. The power of the two-sided test at $\mu = \mu_1$ is then $P(|Y| > z_{1-\alpha/2})$, and this can be evaluated exactly using the command Calc ▶ Probability Distributions ▶ Normal and the distribution function, after writing

$$\begin{aligned} &P\left(|Y| > z_{1-\alpha/2}\right) = P(Y > z_{1-\alpha/2}) + P(Y < -z_{1-\alpha/2}) \\ &= P\left(Z > -\frac{(\mu_1-\mu_0)}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) + P\left(Z < -\frac{(\mu_1-\mu_0)}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right) \end{aligned}$$

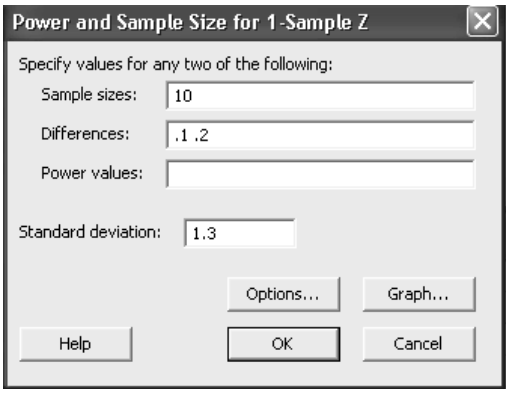
with $Z \sim N(0,1)$.

Alternatively, exact power calculations can be carried out under the assumption of sampling from a normal distribution using the Stat ▶ Power and Sample Size ▶ 1-Sample Z command and filling in the dialog box appropriately. Also,

the minimum sample size required to guarantee a given power at a prescribed difference $|\mu_1 - \mu_0|$ can be obtained using this command. For example, filling in the dialog box for this command as in Display 6.4.1 creates the output

```
Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05 Sigma = 1.3
 Sample
Difference Size    Power
 0.1         10   0.0568057
 0.2         10   0.0775267
```

in the Session window and also produces a graph of the power curve. This gives the power for testing $H_0 : \mu = \mu_0$ versus $H_0 : \mu \neq \mu_0$ at $|\mu_1 - \mu_0| = .1$ and $|\mu_1 - \mu_0| = .2$ when $n = 10$, $\sigma = 1.3$, and $\alpha = .05$. These powers are given by 0.0568057 and 0.0775267, respectively. Clicking on the Options button allows you to choose other alternatives and specify other values of α in the Significance level box.



Display 6.4.1: Dialog box for calculating powers and minimum sample sizes.

If we had instead filled in Power values at .1 and .2 in the dialog box of Display 6.4.1, say as .8 and .9, and had left the Sample sizes box empty, we would have obtained the output

```
Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05 Sigma = 1.3
 Sample Target Actual
  Difference   Size      Power       Power
         0.1     1327     0.8000     0.800160
         0.1     1776     0.9000     0.900039
         0.2      332     0.8000     0.800456
         0.2      444     0.9000     0.900039
```

in the Session window and also a plot of the power curves for each of the different sample sizes. This prescribes the minimum sample sizes $n = 1327$ and $n = 1776$

to obtain the powers .8 and .9, respectively, at the difference .1 and the sample sizes $n = 332$ and $n = 444$ to obtain the powers .8 and .9, respectively, at the difference .2.

This derivation of the power of the two-sided test depended on the sample coming from a normal distribution, as this leads to $\bar{X}$ having an exact normal distribution. In general, however, $\bar{X}$ will be only approximately normal, so the normal calculation for the power is not exact. To assess the effect of the nonnormality, however, we can often simulate sampling from a variety of distributions and estimate the probability $P(|Y| > z_{1-\alpha/2})$. For example, suppose that we want to test $H_0 : \mu = 0$ in a two-sided z test based on a sample of 10, where we estimate σ by the sample standard deviation and we want to evaluate the power at 1. Let us further suppose that we are actually sampling from a uniform distribution on the interval $(-10, 12)$, which indeed has its mean at 1. The simulation given by the session commands

```
MTB > random 1000 c1-c10;
SUBC> uniform -10 12.
MTB > rmean c1-c10 c11
MTB > rstdev c1-c10 c12
MTB > let c13=absolute(c11/(c12/sqrt(10)))
MTB > let c14=c13>1.96
MTB > let k1=mean(c14)
MTB > let k2=sqrt(k1*(1-k1)/n(c14))
MTB > print k1 k2
K1 0.112000
K2 0.00997276
```

estimates the power to be .112, and the standard error of this estimate, as given in K2, is approximately .01. The application determines whether or not the assumption of a uniform distribution makes sense and whether or not this power is indicative of a sensitive test or not.

6.5 The Chi-Square Distribution

If Z is distributed according to the $N(0, 1)$ distribution, then $Y = Z^2$ is distributed according to the Chi-square(1) distribution. If X_1 is distributed Chi-square(k_1) independent of X_2 distributed Chi-square(k_2), then $Y = X_1 + X_2$ is distributed according to the Chi-square($k_1 + k_2$) distribution. There are Minitab commands that assist in carrying out computations for the Chi-square(k) distribution. Note that k is any positive value and is referred to as the *degrees of freedom*.

The values of the density curve for the Chi-square(k) distribution can be obtained using the Calc ▶ Probability _Distributions ▶ Chi-Square command, with k as the Degrees of freedom in the dialog box, or the session command **pdf** with the subcommand **chisquare.** For example, the command

```
MTB > pdf c1 c2;
SUBC> chisquare 4.
```

calculates the value of the Chi-square(4) density curve at each value in C1 and stores these values in C2. This is useful for plotting the density curve. The Calc ▶ Probability Distributions ▶ Chi-Square command, or the session commands **cdf** and **invcdf,** can also be used to obtain values of the Chi-square(k) cumulative distribution function and inverse distribution function, respectively. We use the Calc ▶ Random Data ▶ Chi-Square command, or the session command **random,** to obtain random samples from these distributions.

We will later see applications of the chi-square distribution but we mention one here. In particular, if $x_1, \ldots, x_n$ is a sample from a $N(\mu, \sigma)$ distribution, then $(n-1)\, s^2/\sigma^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / \sigma^2$ is known to follow a Chi-square($n-1$) distribution, and this fact is used as a basis for inference about σ (confidence intervals and tests of significance). Because of the nonrobustness of these inferences to small deviations from normality, these inferences are not recommended.

6.6 Exercises

If your version of Minitab places restrictions such that the value of the simulation sample size N requested in these problems is not feasible, then substitute a more appropriate value. Be aware, however, that the accuracy of your results is dependent on how large N is.

1. Suppose we obtain the following sample from a $N(\mu, 2.3)$ distribution.

−1.3	−2.5	−2.5	−0.9	1.8	−2.9	−3.0	1.7	−0.1	−4.8

   Use the Stat ▶ Basic Statistics ▶ 1- Sample Z command to compute 90%, 95%, and 99% confidence intervals for μ.

2. For the data in Exercise II.6.1, use the Stat ▶ Basic Statistics ▶ 1- Sample Z command to test the null hypothesis $H_0 : \mu = 0$ in a two-sided test. Evaluate the power of the test with level $\alpha = .05$ at $\mu = 1$. Repeat these calculations but this time test the null hypothesis $H_0 : \mu \leq 0$.

3. Simulate $N = 1000$ samples of size 5 from the $N(1, 2)$ distribution, and calculate the proportion of .90 z confidence intervals for the mean that cover the true value $\mu = 1$.

4. Simulate $N = 1000$ samples of size 10 from the uniform distribution on (0,1), and calculate the proportion of .90 z confidence intervals for the mean that cover the true value $\mu = .5$. Use $\sigma = 1/\sqrt{12}$.

5. Simulate $N = 1000$ samples of size 10 from the Exponential(1) distribution (see Exercise II.4.7), and calculate the proportion of .95 z confidence intervals for the mean that cover the true value $\mu = 1$. Use $\sigma = 1$.

6. The density curve for the Student(1) distribution takes the form

$$\frac{1}{\pi}\frac{1}{1+x^2}$$

for $-\infty < x < \infty$. This special case is called the *Cauchy* distribution. Plot this density curve in the range $(-20, 20)$ using an increment of .1. Simulate $N = 1000$ samples of size 5 from the Student(1) distribution (see Exercise II.4.12), and calculate the proportion of .90 confidence intervals for the mean, using the sample standard deviation for σ, that cover the value $\mu = 0$. It is possible to obtain very bad approximations in this example because the central limit theorem does not apply to this distribution. In fact, it does not have a mean.

7. Suppose we are testing $H_0 : \mu = 3$ versus $H_0 : \mu \neq 3$ when we are sampling from a $N(\mu, \sigma)$ distribution with $\sigma = 2.1$ and the sample size is $n = 20$. If we use the critical value $\alpha = .01$, determine the power of this test at $\mu = 4$.

8. Suppose we are testing $H_0 : \mu = 3$ versus $H_0 : \mu > 3$ when we are sampling from a $N(\mu, \sigma)$ distribution with $\sigma = 2.1$. If we use the critical value $\alpha = .01$, determine the minimum sample size so that the power of this test at $\mu = 4$ is .99.

9. The uniform distribution on the interval (a, b) has mean $\mu = (a + b)/2$ and standard deviation $\sigma = ((b - a)^2/12)^{1/2}$. Calculate the power at $\mu = 1$ of the two-sided z test at level $\alpha = .95$ for testing $H_0 : \mu = 0$ when the sample size is $n = 10$, σ is the standard deviation of a uniform distribution on $(-10, 12)$, and we are sampling from a normal distribution.

10. Suppose that we are testing $H_0 : \mu = 0$ in a two-sided test based on a sample of 3. Approximate the power of the z test at level $\alpha = .1$ at $\mu = 5$ when we are sampling from the distribution of $Y = 5 + W$, where W follows a Student(6) distribution (see Exercise II.4.12) and we use the sample standard deviation to estimate σ. Note that the mean of the distribution of Y is 5.

Chapter 7

Inference for Distributions

New Minitab commands discussed in this chapter

Calc ▶ Probability Distributions ▶ F
Calc ▶ Probability Distributions ▶ t
Calc ▶ Random Data ▶ F
Calc ▶ Random Data ▶ t
Power and Sample Size ▶ 1-Sample t
Power and Sample Size ▶ 2-Sample t
Stat ▶ Basic Statistics ▶ 1-Sample t
Stat ▶ Basic Statistics ▶ 2-Sample t
Stat ▶ Nonparametrics ▶ 1-Sample Sign

7.1 The Student Distribution

If Z is distributed $N(0,1)$ independent of X distributed Chi-square(k) (see II.6.5), then $T = Z/\sqrt{X/k}$ is distributed according to the Student(k) distribution. The value k is referred to as the *degrees of freedom* of the Student distribution. There are Minitab commands that assist in carrying out computations for this distribution.

The values of the density curve, distribution function, and inverse distribution function for the Student(k) distribution can be obtained using the Calc ▶ Probability Distributions ▶ t command with k as the Degrees of freedom. Alternatively, we can use the session commands **pdf, cdf,** and **invcdf** with the **student** subcommand. For example, the command

```
MTB > pdf c1 c2;
SUBC> student 4.
```

calculates the value of the Student(4) density curve at each value in C1 and stores these values in C2. This is useful for plotting the density curve. To

generate from this distribution we use the command Calc ▶ Random Data ▶ t again with k as the Degrees of freedom or use the session command **random** with the **student** subcommand.

7.2 t Confidence Intervals

When sampling from the $N(\mu, \sigma)$ distribution with μ and σ unknown, an exact $1-\alpha$ confidence interval for μ based on the sample $x_1, \ldots, x_n$ is given by $\bar{x} \pm t^* s/\sqrt{n}$, where t^* is the $1-\alpha/2$ percentile of the Student$(n-1)$ distribution. These intervals can be obtained using the Stat ▶ Basic Statistics ▶ 1- Sample t command.

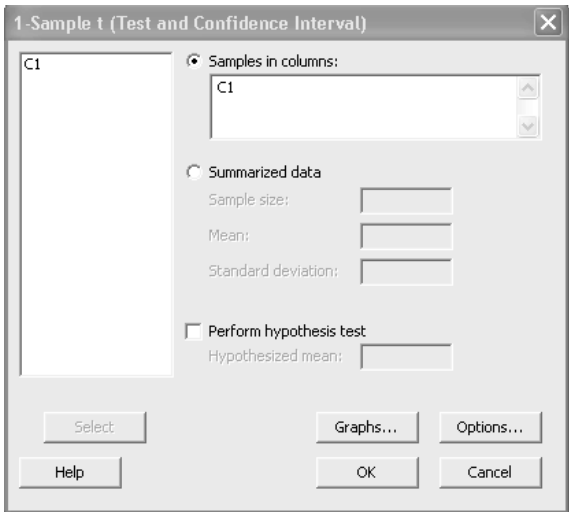
For example, suppose that we have the following sample of $n = 10$ in C1.

0.44	4.19	0.22	4.23	1.46
3.98	2.29	1.79	6.09	3.04

Then the Stat ▶ Basic Statistics ▶ 1- Sample t command, with the dialog box as in Display 7.2.1, produces the output

```
Variable   N    Mean    StDev   SE Mean        95% CI
C1        10   2.773   1.872    0.592     (1.434, 4.112)
```

in the Session window. This computes a 95% confidence interval for μ as $(1.43372, 4.11228)$. To change the confidence level, click on the Options button and fill in the subsequent dialog box appropriately.



Display 7.2.1: Dialog box for producing t confidence intervals.

The general syntax of the corresponding session command **tinterval** is

tinterval V $E_1 \ldots E_m$

where V is the confidence level and is any value between 1 and 99.99 and E_1, ..., E_m are columns of data. A V% confidence interval is produced for each column specified. If no value is specified for V, the default value is 95%.

7.3 t Tests

The Stat ▶ Basic Statistics ▶ 1-Sample t command is used when we have a sample $x_1, \ldots, x_n$ from a normal distribution with unknown mean μ and standard deviation σ and we want to test the hypothesis that the unknown mean equals a value μ_0. The test is based on computing a P-value using the observed value of

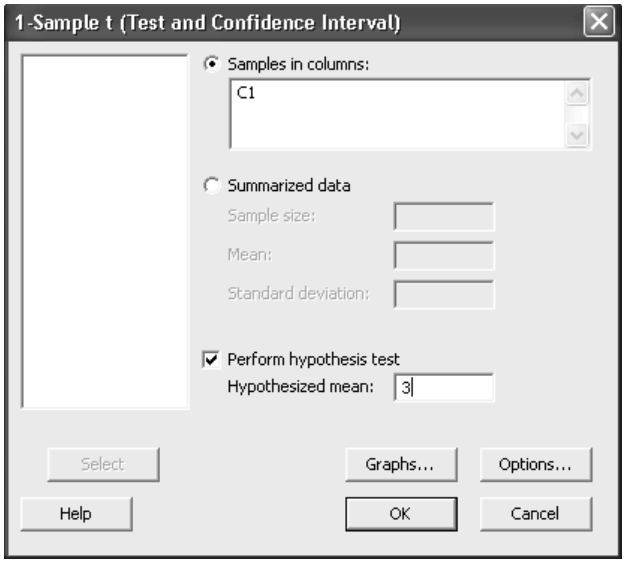
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and the Student$(n-1)$ distribution.

For example, suppose we want to test $H_0 : \mu = 3$ for the data presented in Section 7.2. Then the Stat ▶ Basic Statistics ▶ 1- Sample t command, with the dialog box as in Display 7.3.1, produces the output

```
Test of mu = 3 vs not = 3
Variable  N    Mean     StDev   SE Mean
C1        10   2.773    1.872   0.592
       95% CI          T       P
   (1.434, 4.112)   -0.38   0.710
```

so we have the P-value as 0.710 and we have no evidence against $H_0 : \mu = 3$. To assess other hypotheses click on the Options button and fill in the subsequent dialog box appropriately.



Display 7.3.1: First dialog box for a test of hypothesis using the t statistic.

The general syntax of the corresponding session command **ttest** is

ttest V $E_1 \ldots E_m$

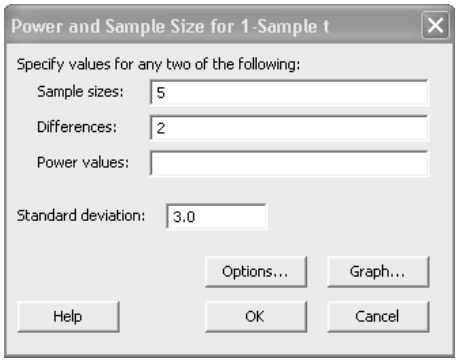
where V is the hypothesized value to be tested and E_1, ..., E_m are columns of data. If no value is specified for V, the default is 0. A test of the hypothesis is carried out for each column. Also, the **alternative** subcommand is available and works just as with the **ztest** command.

Note that the Stat ▶ Basic Statistics ▶ 1-Sample t command can also be used to carry out t tests for the difference of two means in a matched pairs design. For this, store the difference of the measurements in a column and apply Stat ▶ Basic Statistics ▶ 1-Sample t to that column as shown previously.

Exact power calculations can be carried under the assumption of sampling from a normal distribution using Power and Sample Size ▶ 1-Sample t and filling in the dialog box appropriately. Further, the minimum sample size required to guarantee a given power at a prescribed difference $|\mu_1 - \mu_0|$ and standard deviation σ can be obtained using this command. For example, using this command with the dialog box as in Display 7.3.2, we obtain the output

```
Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05 Sigma = 3
 Sample
Difference   Size     Power
    2          5    0.2113
```

in the Session window together with a plot of the power curve for this test. This gives the exact power of the two-sided t test when $n = 5, |\mu_1 - \mu_0| = 2, \sigma = 3.0$, and $\alpha = .05$ as .2113. The Options button can be used to compute power for one-sided tests.



Display 7.3.2: Dialog box for determining power and minimum sample sizes when using the one-sample t test.

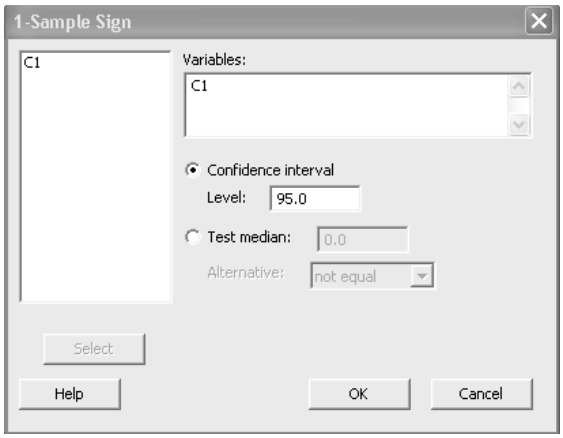
7.4 The Sign Test

Sometimes we cannot sensibly assume normality or transform to normality or don't have a large sample so that there is a central limit theorem effect. In such a case, we attempt to use *distribution free* or *nonparametric* methods. The testing method based on the *sign test statistic* for the median is one of these.

For example, suppose we have the data of Section 7.2 stored in column C1. Then the Stat ▶ Nonparametrics ▶ 1-Sample Sign command produces the dialog box given in Display 7.4.1. Here we have filled in the Confidence interval button, and in the Level box we have requested a .95 confidence interval for the median. The following output is obtained.

```
Sign confidence interval for median
                                  Confidence
                  Achieved          Interval
     N  Median  Confidence  Lower  Upper   Position
C1  10   2.665    0.8906    1.460  4.190      3
                  0.9500    1.111  4.204     NLI
                  0.9785    0.440  4.230      2
```

As the distribution of the sign statistic is discrete, in general the exact confidence cannot be attained, so Minitab records the confidence intervals with confidence level just smaller and just greater than the confidence level requested and also records a middle interval obtained by interpolation.



Display 7.4.1: Dialog box for the sign test and the sign confidence interval.

If instead we fill in the Test median button and enter 4.0 for the null hypothesis with the Alternative not equal, we obtain the output

```
Sign test of median = 4.000 versus not = 4.000
     N  Below  Equal  Above   P      Median
C1  10    7      0      3   0.3438   2.665
```

which gives the P-value as 0.3438 for assessing the hypothesis that the median of the population distribution equals 4.0. Also, the sample median of 2.665 is recorded.

Note that the Stat ▶ Nonparametrics ▶ 1-Sample Sign command can also be used to construct confidence intervals and carry out tests for the median of a difference in a matched pairs design. For this, store the difference of the measurements in a column and apply the command to that column.

The corresponding session commands are **sinterval,** for the sign confidence interval and **stest,** for the sign test. The general syntax of the **sinterval** command is

sinterval V $E_1 \ldots E_m$

where V is the confidence level, and is any value between 1 and 99.99, and E_1, ..., E_m are columns of data. A V% confidence interval is produced for each column specified. If no value is specified for V, then the default value is 95%. The general syntax of the **stest** command is

stest V $E_1 \ldots E_m$

where V is the hypothesized value to be tested and E_1, ..., E_m are columns of data. If no value is specified for V, the default is 0. A test of the hypothesis is carried out for each column. The **alternative** subcommand is also available for one-sided tests.

7.5 Comparing Two Samples

If we have independent samples $x_{11}, \ldots x_{1n_1}$ from the $N(\mu_1, \sigma_1)$ distribution and $x_{12}, \ldots x_{1n_2}$ from the $N(\mu_2, \sigma_2)$ distribution, where σ_1 and σ_2 are known, we can base inferences about the difference of the means $\mu_1 - \mu_2$ on the z statistic given by

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Under these assumptions, z has an $N(0,1)$ distribution. Therefore, a $1 - \alpha$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{x}_1 - \bar{x}_2 \pm \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} z^*$$

where z^* is the $1 - \alpha/2$ percentile of the $N(0,1)$ distribution. We can test $H_0 : \mu = \mu_0$ against the alternative $H_a : \mu \neq \mu_0$ by computing the P-value $P(|Z| > |z_0|) = 2P(Z > z_0)$, where Z is distributed $N(0,1)$ and z_0 is the observed value of the z statistic. These inferences are also appropriate without normality, provided n_1 and n_2 are large and we have the values σ_1 and σ_2 or good estimates. These inferences are easily carried out using Minitab commands we have already discussed.

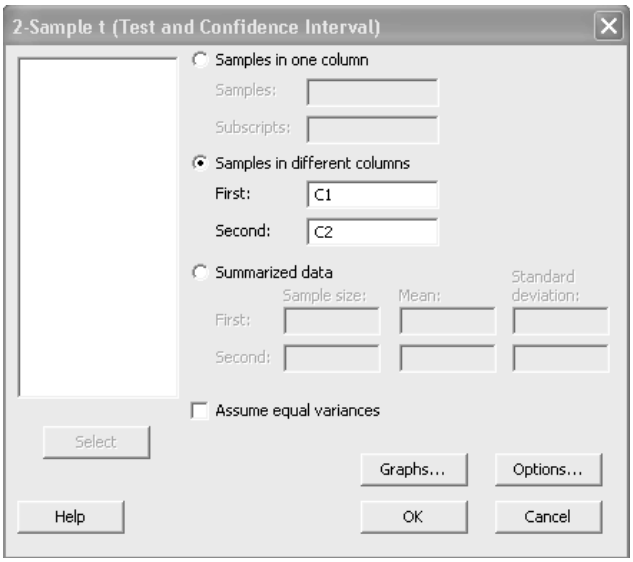
In general, however, we will not have available suitable values of σ_1 and σ_2 or large samples and will have to use the two-sample analogs of the single-sample t procedures just discussed. This is acceptable, provided, of course, that we

have checked for and agreed that it is reasonable to assume that both samples are from normal distributions. These procedures are based on the two-sample t statistic given by

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where we have replaced the population standard deviations by their sample estimates. The exact distribution of this statistic does not have a convenient form, but, of course, we can always simulate its distribution. Actually, it is typical to use an approximation to the distribution of this statistic based on a Student distribution. Use Help to get more details on this.

The Stat ▶ Basic Statistics ▶ 2-Sample t command is available for computing inference procedures based on t, using a dialog box as in Display 7.5.1.



Display 7.5.1: Dialog box for two sample problems based on the two-sample t statistic.

For example, suppose that we have the following values for two samples,

Sample 1	7	−4	18	17	−3	−5	1	10	11	−2	
Sample 2	−1	12	−1	−3	3	−5	5	2	−11	−1	−3

with Sample 1 in C1 and Sample 2 in C2. The Stat ▶ Basic Statistics ▶ 2-Sample t command with the dialog box as in Display 7.5.1 produces the output

```
Two-sample T for C1 vs C2
    N   Mean   StDev   SE Mean
C1 10   5.00    8.74    2.8
C2 11  -0.27    5.90    1.8
Difference = mu (C1) - mu (C2)
Estimate for difference:  5.27
95% CI for difference:  (-1.74, 12.28)
T-Test of difference = 0 (vs not =):  T-Value = 1.60
P-Value = 0.130 DF = 15
```

in the Session window. This gives a 95% confidence interval for the difference in the means $\mu_1 - \mu_2$ as $(-1.74, 12.28)$ and calculates the P-value .130 for the test of $H_0 : \mu_1 - \mu_2 = 0$ versus the alternative $H_a : \mu_1 - \mu_2 \neq 0$. In this case, we do not reject H_0.

Notice we have selected the Samples in different columns radio button, as this is how we have stored our data. Alternatively, we can store all the actual measurements in a single column with a second column providing an index of the sample to which the observation belongs. Clicking on the Options button of the dialog box of Display 7.5.1 produces a dialog box where we can prescribe a different value for the confidence level, the difference between the means that we wish to test for, and the type of hypothesis.

Notice also that, in the dialog box of Display 7.5.1, we have left the box Assume equal variances unchecked. This box is checked only when we feel that we can assume that $\sigma_1 = \sigma_2 = \sigma$ and want to pool both samples together to estimate the common σ. Pooling is usually unnecessary and is not recommended.

Power calculations can be carried out under the assumption of sampling from a normal distribution using Power and Sample Size ▶ 2-Sample t and filling in the dialog box appropriately, although this requires the assumption of a common population standard deviation σ. Further, the minimum sample size required to guarantee a given power at a prescribed difference $|\mu_1 - \mu_2|$, and assuming a common standard deviation σ, can be obtained using this command. This command works the same as the one sample case.

There are two corresponding session commands—**twosample** and **twot.** Each of these commands computes confidence intervals for the difference of the means and computes P-values for tests of significance concerning the difference of means. The only difference between these commands is that with **twosample** the two samples are in individual columns, while with **twot** the samples are in a single column with subscripts indicating group membership in a second column. The general syntax of the **twosample** command is

twosample V E_1 E_2

where V is the confidence level and is any value between 1 and 99.99 and E_1, E_2 are columns of data containing the two samples. The general syntax of the **twot** command is

twot V E_1 E_2

where V is the confidence level and is any value between 1 and 99.99 and E_1, E_2 are columns of data with E_1 containing the samples and E_2 containing the subscripts.

The **alternative** subcommand is available with both **twosample** and **twot** if we wish to conduct one-sided tests. Also, the subcommand **pooled** is available if we feel we can assume that $\sigma_1 = \sigma_2 = \sigma$ and want to pool both samples together to estimate the common σ.

7.6 The F Distribution

If X_1 is distributed Chi-square(k_1) independent of X_2 distributed Chi-square(k_2), then

$$F = \frac{X_1/k_1}{X_2/k_2}$$

is distributed according to the $F(k_1, k_2)$ distribution. The value k_1 is called the *numerator degrees of freedom* and the value k_2 is called the *denominator degrees of freedom*. There are Minitab commands that assist in carrying out computations for this distribution.

The values of the density curve for the $F(k_1, k_2)$ distribution can be obtained using the Calc ▶ Probability Distributions ▶ F command, with k_1 specified as the Numerator degrees of freedom and k_2 specified as the Denominator degrees of freedom in the dialog box. For example, this command with the dialog box as in Display 7.6.1 produces the output

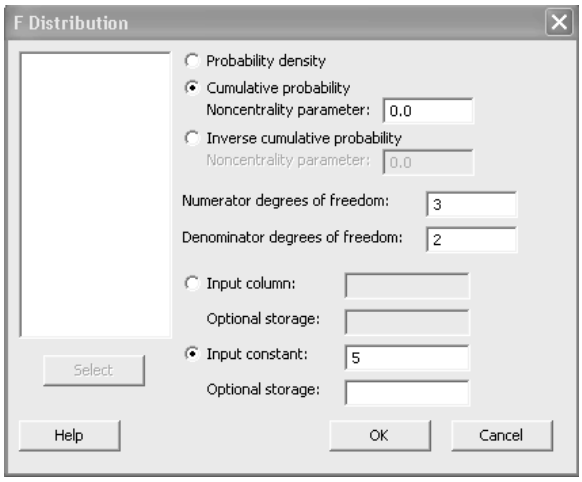
```
      x          P( X <= x )
 5.0000            0.828826
```

in the Session window. This calculates the value of the $F(3, 2)$ distribution function at 5 as .8288. Alternatively, you can use the session commands **pdf, cdf,** and **invcdf** with the **F** subcommand. The Calc ▶ Random Data ▶ F command and the session command **random** with the **F** subcommand can be used to obtain random samples from the $F(k_1, k_2)$ distribution.

There are a number of applications of the F-distribution. In particular, if $x_{11}, \ldots x_{1n_1}$ is a sample from the $N(\mu_1, \sigma_1)$ distribution and $x_{12}, \ldots x_{1n_2}$ a sample from the $N(\mu_2, \sigma_2)$ distribution, then

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

is known to follow an $F(n_1 - 1, n_2 - 1)$ distribution. This fact is used as a basis for inference about the ratio σ_1/σ_2, i.e., confidence intervals and tests of significance and, in particular, testing for equality of variances between the samples. Because of the nonrobustness of these inferences to small deviations from normality, these inferences are not usually recommended.



Display 7.6.1: Dialog box for probability calculations for the $F(k_1, k_2)$ distribution.

7.7 Exercises

If your version of Minitab places restrictions such that the value of the simulation sample size N requested in these problems is not feasible, then substitute a more appropriate value. Be aware, however, that the accuracy of your results is dependent on how large N is.

1. Plot the Student(k) density curve for $k = 1, 2, 10, 30$ and the $N(0, 1)$ density curve on the interval $(-10, 10)$ using an increment of .1 and compare the plots.

2. Make a table of the values of the cumulative distribution function of the Student(k) distribution for $k = 1, 2, 10, 30$ and the $N(0, 1)$ distribution at the points $-10, -5, -3, -1, 0, 1, 3, 5, 10$. Comment on the values.

3. Make a table of the values of the inverse cumulative distribution function of the Student(k) distribution for $k = 1, 2, 10, 30$ and the $N(0, 1)$ distribution at the points $.0001, .001, .01, .1, .25, .5$. Comment on the values.

4. Simulate $N = 1000$ values from Z distributed $N(0, 1)$ and X distributed Chi-square(3) and plot a histogram of $T = Z/\sqrt{X/3}$ using the cutpoints $-10, -9, \ldots, 9, 10$. Generate a sample of $N = 1000$ values directly from the Student(3) distribution, plot a histogram with the same cutpoints, and compare the two histograms.

5. Carry out a simulation with $N = 1000$ to verify that the 95% confidence interval based on the t statistic covers the true value of the mean 95% of the time when taking samples of size 5 from the $N(4, 2)$ distribution.

6. Generate a sample of 50 from the $N(10, 2)$ distribution. Compare the 95% confidence intervals obtained via the Stat ▶ Basic Statistics ▶ 1-Sample

t and Stat ▶ Basic Statistics ▶ 1- Sample Z commands using the sample standard deviation as an estimate of σ.

7. Calculate the power of the t test at $\mu_1 = 1, \sigma_1 = 2$ for testing $H_0 : \mu = 0$ versus the alternative $H_a : \mu \neq 0$ at level $\alpha = .05$, based on a sample of 5 from the normal distribution.

8. Simulate the power of the two sample t test at $\mu_1 = 1, \sigma_1 = 2, \mu_2 = 2, \sigma_1 = 3$ for testing $H_0 : \mu_1 - \mu_2 = 0$ versus the alternative $H_a : \mu_1 - \mu_2 \neq 0$ at level $\alpha = .05$, based on a sample of 5 from the $N(\mu_1, \sigma_1)$ distribution and a sample of size 8 from the $N(\mu_2, \sigma_2)$ distribution. Use the conservative rule when choosing the degrees of freedom for the approximate test, i.e., the smaller of $n_1 - 1$ and $n_2 - 1$.

9. If Z is distributed $N(\mu, 1)$ and X is distributed Chi-square(k) independent of Z, then
$$Y = \frac{Z}{\sqrt{X/k}}$$
is distributed according to a *noncentral* Student(k) distribution with noncentrality μ. Simulate samples of $N = 1000$ from this distribution with $k = 5$ and $\mu = 0, 1, 5, 10$. Plot the samples in histograms with cutpoints $-20, -19, \ldots, 19, 20$ and compare these plots.

10. If X_1 is distributed Chi-square(k_1) independently of X_2, which is distributed $N(\delta, 1)$, then the random variable $Y = X_1 + X_2^2$ is distributed according to a *noncentral* Chi-square($k + 1$) distribution with noncentrality $\lambda = \delta^2$. Generate samples of $n = 1000$ from this distribution with $k = 2$ and $\lambda = 0, 1, 5, 10$. Plot histograms of these samples with the cut-points 0,1, ..., 200. Comment on the appearance of these histograms.

11. If X_1 is distributed *noncentral* Chi-square(k_1) with non-centrality λ independently of X_2, which is distributed Chi-square(k_2), then the random variable
$$Y = \frac{X_1/k_1}{X_2/k_2}$$
is distributed according to a *noncentral* $F(k_1, k_2)$ distribution with noncentrality λ. Generate samples of $n = 1000$ from this distribution with $k_1 = 2, k_2 = 3$, and $\lambda = 0, 1, 5, 10$. Plot histograms of these samples with the cut-points 0,1, ..., 200. Comment on the appearance of these histograms.

Chapter 8

Inference for Proportions

New Minitab commands discussed in this chapter

Power and Sample Size ▶ 1 Proportion
Power and Sample Size ▶ 2 Proportions
Stat ▶ Basic Statistics ▶ 1 Proportion
Stat ▶ Basic Statistics ▶ 2 Proportions

This chapter is concerned with inference methods for a proportion p and for the comparison of two proportions p_1 and p_2. Proportions arise from measuring a binary-valued categorical variable on population elements, such as gender in human populations. For example, p might be the proportion of females in a given population, or we might want to compare the proportion p_1 of females in population 1 with the proportion p_2 of females in population 2. The need for inference arises as we base our conclusions about the values of these proportions on samples from the populations rather than measuring every element in the population. For convenience, we will denote the values assumed by the binary categorical variables as 1 and 0, where 1 indicates the presence of a characteristic and 0 indicates its absence.

8.1 Inference for a Single Proportion

Suppose that $x_1, \ldots, x_n$ is a sample from a population where the variable is measuring the presence or absence of some trait by a 1 or 0, respectively. Let $\hat{p}$ be the proportion of 1's in the sample. This is the estimate of the true proportion p. For example, the sample could arise from coin tossing, where 1 denotes heads and 0 tails and $\hat{p}$ is the proportion of heads, while p is the probability of heads. If the population we are sampling from is finite, then, strictly speaking, the sample elements are not independent. But if the population size is large relative to the sample size n, then independence is a reasonable approximation, and this is

necessary for the methods of this chapter. So we will consider $x_1, \ldots, x_n$ as a sample from the Bernoulli(p) distribution.

The standard error of the estimate $\hat{p}$ is $\sqrt{\hat{p}(1-\hat{p})/n}$, and because $\hat{p}$ is an average, the central limit theorem gives that

$$z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

is approximately $N(0,1)$ for large n. This leads to the approximate $1-\alpha$ confidence interval given by $\hat{p} \pm \sqrt{\hat{p}(1-\hat{p})/n}z^*$, where z^* is the $1-\alpha/2$ percentile of the $N(0,1)$ distribution. To test a null hypothesis $H_0 : p = p_0$, we make use of the fact that under the null hypothesis the statistic

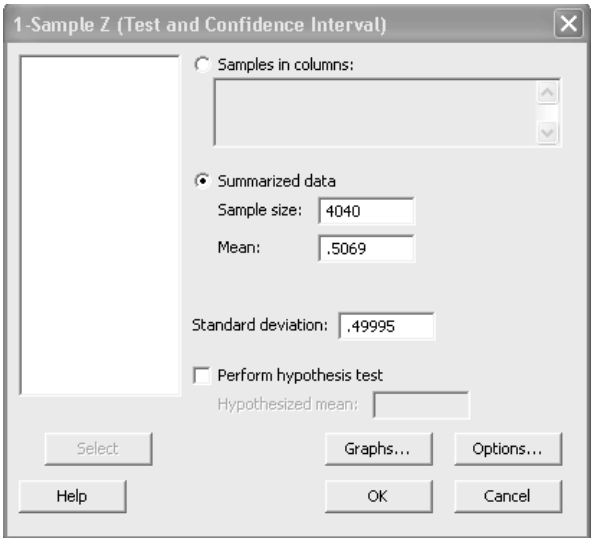
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

is approximately $N(0,1)$. To test $H_0 : p = p_0$ versus $H_a : p \neq p_0$, we compute $P(|Z| > |z|) = 2P(Z > |z|)$, where Z is distributed $N(0,1)$.

For example, suppose that a coin was tossed $n = 4040$ times and the observed proportion of heads is $\bar{x} = 2048/4040 = .5069$. Then we have that $\sqrt{.5069(1-.5069)} = 0.49995$ and, using Stat ▶ Basic Statistics ▶ 1-Sample Z and the dialog box in Display 8.1.1, we obtain the output

```
   N      Mean       SE Mean          95% CI
4040   0.50690     0.00787    (0.49148, 0.52232)
```

which provides an approximate .95-confidence interval for p.

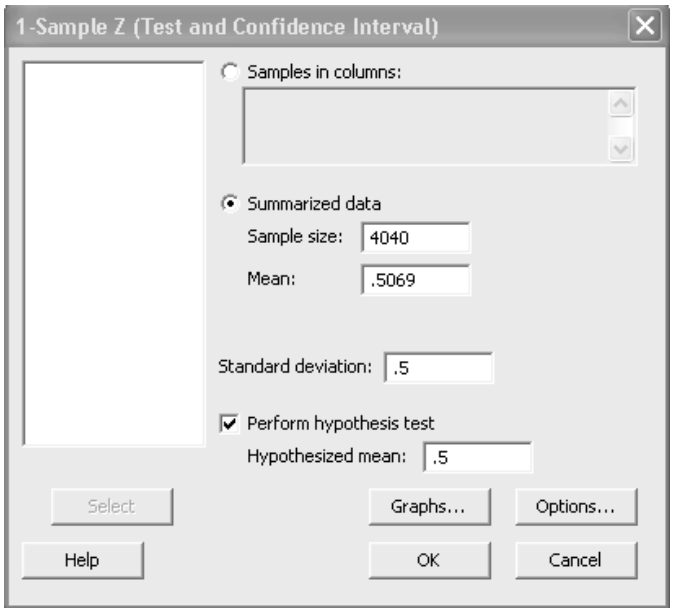


Display 8.1.1: Dialog box for obtaining confidence intervals.

Similarly, if we want to assess the hypothesis $H_0 : p = .5$, then $\sqrt{.5(1-.5)} = 0.5$ and Stat ▶ Basic Statistics ▶ 1-Sample Z and the dialog box in Display 8.1.2 leads to

```
Test of mu = 0.5 vs not = 0.5
The assumed standard deviation = 0.5
  N     Mean     SE Mean        95% CI          Z     P
4040  0.50690    0.00787  (0.49148, 0.52232)  0.88  0.380
```

which gives the P-value as 0.380 and so we have no evidence against $H_0 : p = .5$. If we wish to use other confidence levels or test other hypotheses, then these options are available using the Options button in the dialog box.

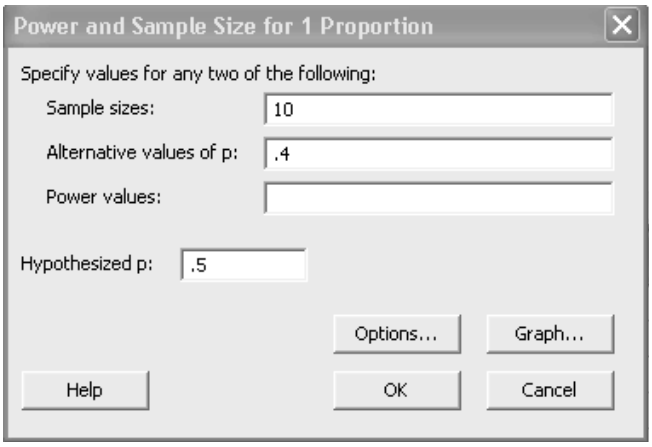


Display 8.1.2: Dialog box for obtaining P-values.

Note that the estimate and confidence intervals recorded by the software are not those based on the *Wilson estimate*. To obtain the Wilson estimate and the associated confidence interval, we must add four data values to the data set—two heads (or successes) and two tails (or failures). So in this case, implementing the above command with the number of trials equal to 4044 and the number of successes equal to 2050 will produce the inferences based on the Wilson estimate.

Power calculations and minimum sample sizes to achieve a prescribed power can be obtained using Power and Sample Size ▶ 1 Proportion. For example, suppose we want to compute the power of the test for $H_0 : p = .5$ versus $H_a : p \neq .5$ at level $\alpha = .05$ at $n = 10, p = .4$. This command, with the dialog

box as in Display 8.1.3,



Display 8.1.3: Dialog box for power calculations for test of a single proportion.

produces the output

```
Testing proportion = 0.5 (versus not = 0.5)
Alpha = 0.05
    Alternative      Sample
    Proportion        Size       Power
      0.4              10      0.0918014
```

which calculates this power as .0918014. So the test is not very powerful. By contrast, at $n = 100, p = .4$ the power is .51633.

8.2 Inference for Two Proportions

Suppose that $x_{11}, \ldots, x_{n_1 1}$ is a sample from population 1 and $x_{12}, \ldots, x_{n_2 2}$ is a sample from population 2, where the variable is measuring the presence or absence of some trait by a 1 or 0, respectively. We assume then that we have a sample of n_1 from the Bernoulli(p_1) distribution and a sample of n_2 from the Bernoulli(p_2) distribution. Suppose that we want to make inferences about the difference in the proportions $p_1 - p_2$. Let $\hat{p}_i$ be the proportion of 1's in the ith sample.

The central limit theorem gives that

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

is approximately $N(0,1)$ for large n_1 and n_2. This leads to the approximate $1 - \alpha$ confidence interval given by

$$\hat{p}_1 - \hat{p}_2 \pm \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\, z^*$$

where z^* is the $1 - \alpha/2$ percentile of the $N(0,1)$ distribution. The Wilson estimate and its corresponding confidence interval are obtained by adding four data values to the data set—one success and one failure to each sample—so that the ith sample size becomes $n_i + 2$ and the ith sample estimate becomes $(n_i\hat{p}_i + 1)/(n_i + 2)$. The above formula for the confidence interval applied with these changes then gives the interval based on the Wilson estimates.

To test a null hypothesis $H_0 : p_1 = p_2$ we use the fact that under the null hypothesis the statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is approximately $N(0,1)$ for large n_1 and n_2, where

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

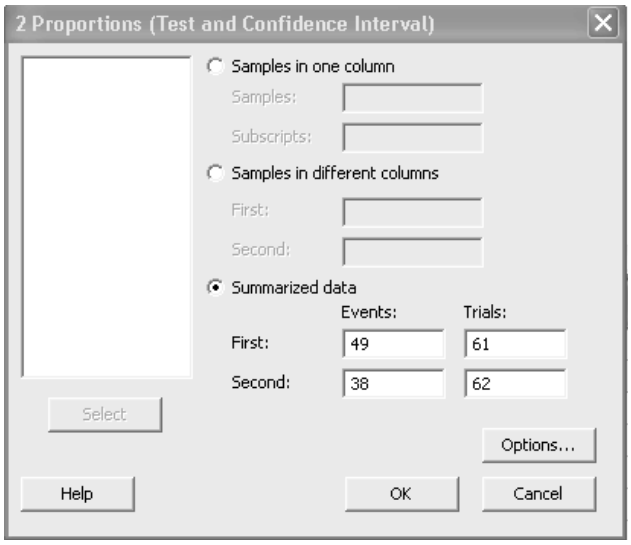
is the estimate of the common value of the proportion when the null hypothesis is true. To test $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$ we compute $P(|Z| > |z|) = 2P(Z > |z|)$ where Z is distributed $N(0,1)$.

For example, suppose that we want to test $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$ when $n_1 = 61, \hat{p}_1 = .803 = 49/61, n_2 = 62$, $\hat{p}_2 = .613 = 38/62$. The command Stat ▶ Basic Statistics ▶ 2 Proportions with the dialog box as in Display 8.2.1 produces the output

```
    Sample    X     N     Sample p
      1      49    61     0.803279
      2      38    62     0.612903

Difference = p (1) - p (2)

Estimate for difference:  0.190375

95% CI for difference:  (0.0333680, 0.347383)

Test for difference = 0 (vs not = 0):  Z = 2.38 P-Value = 0.017
```

in the Session window. The P-value is .017, so we would definitely reject. A 95% confidence interval for $p_1 - p_2$ is given by $(0.0333680, 0.347383)$. If other tests or confidence intervals are required, then these are available via the Options button. The Wilson estimates and associated confidence interval are obtained from the software by modifying the data as indicated above.

Power calculations and minimum sample sizes to achieve a prescribed power can be obtained using Power and Sample Size ▶ 2 Proportions.



Display 8.2.1: Dialog box for inferences comparing two proportions.

8.3 Exercises

Don't forget to quote standard errors for any approximate probabilities you quote in the following problems.

1. Carry out a simulation with the Binomial$(40, .3)$ distribution to assess the coverage of the 95% confidence interval for a single proportion.

2. The accuracy of a confidence interval procedure can be assessed by computing *probabilities of covering false values.* Approximate the probabilities of covering the values .1, .2, ..., .9 for the 95% confidence interval for a single proportion when sampling from the Binomial$(20, .5)$ distribution.

3. Calculate the power of the two-sided test for testing $H_0 : p = .5$ at level $\alpha = .05$ at the points $n = 100, p = .1, ..., 9$ and plot the power curve.

4. Carry out a simulation with the Binomial$(40, .3)$ and the Binomial$(50, .4)$ distribution to assess the coverage of the 95% confidence interval for a difference of proportions.

5. Calculate the power of the two-sided test for testing $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$ at level $\alpha = .05$ at $n_1 = 40, p_1 = .3, n_2 = 50, p_2 = .1, ..., 9$ and plot the power curve.

Chapter 9

Inference for Two-Way Tables

New Minitab commands discussed in this chapter

Stat ▶ Tables ▶ Chi-Square Goodness-of-Fit Test (One variable)
Stat ▶ Tables ▶ Chi-Square Test (Two-Way Table in Worksheet)
Stat ▶ Tables ▶ Cross Tabulation and Chi-Square

In this chapter, inference methods are discussed for comparing the distributions of a categorical variable for a number of populations and for looking for relationships among a number of categorical variables defined on a single population. The *chi-square test* is the basic inferential tool, and this is implemented in Minitab via the Stat ▶ Tables ▶ Cross Tabulation and Chi-Square command, if the data is in the form of raw incidence data, or the Stat ▶ Tables ▶ Chi-Square Test command, if the data comes in the form of counts.

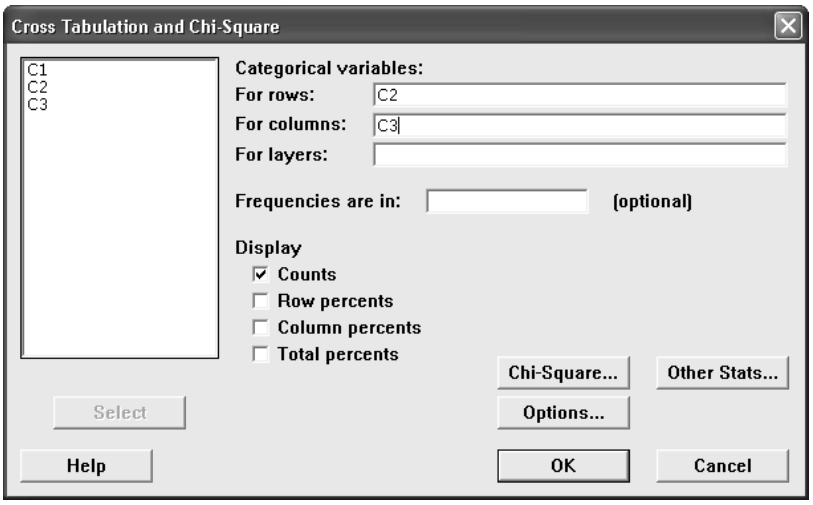
9.1 Tabulating and Plotting

The relationship between two categorical variables is typically assessed by cross-tabulating the variables in a table. For this, the Stat ▶ Tables ▶ Cross Tabulation and Chi-Square command is available. We illustrate using an example where each categorical variable takes two values. Of course, each variable can take a number of values, and this need not be the same for each categorical variable.

Suppose that we have collected data on courses being taken by students and have recorded a 1 in C2 if the student is taking Statistics and a 0 if not. If the student is taking Calculus, a 1 is recorded in C3 and a 0 otherwise. Also, we have recorded the student number in C1. These data for 10 students follow.

Row	C1	C2	C3
1	12389	1	0
2	97658	1	0
3	53546	0	1
4	55542	0	1
5	11223	1	1
6	77788	0	0
7	44567	1	1
8	32156	1	0
9	33456	0	1
10	67945	0	1

We cross-tabulate the data in C2 and C3 using the Stat ▶ Tables ▶ Cross Tabulation and Chi-Square command and the dialog box shown in Display 9.1.1.



Display 9.1.1: Dialog box for producing tables.

This produces the output

```
Rows:  C2     Columns:  C3

          0        1        All

0         1        4          5
1         3        2          5
All       4        6         10

  Cell Contents:                Count
```

in the Session window that reveals there is 1 student taking neither Statistics nor Calculus, 4 students taking Calculus but not Statistics, 3 students taking Statistics but not Calculus, and 2 students taking both subjects. The row

marginal totals are produced on the right, and the column marginal totals are produced below the table. We have chosen the cell entries in the table to be frequencies (counts), but we can see from Display 9.1.1 that there are other choices. For example, if we had checked the Total percents box instead, we obtain the output

```
Rows:  C2       Columns:  C3

            0        1         All

0       10.00    40.00      50.00
1       30.00    20.00      50.00
All     40.00    60.00     100.00

Cell Contents:                      % of Total
```

where each entry is the percentage that cell represents of the total number of observations used to form the table. Of course, we can ask for more than just one of these cell statistics to be produced in a table.

To examine the relationship between the two variables, we compare the conditional distributions given row, by checking the Row percents box, or the conditional distributions given column, by checking the Column percents box. For example, choosing to calculate row percents gives us the table

```
Rows:  C2         Columns:  C3

           0        1       All

0      20.00    80.00    100.00
1      60.00    40.00    100.00
All    40.00    60.00    100.00

Cell Contents:                  % of Row
```

that gives the row distributions as 20%, 80% for the first row and 60%, 40% for the second row. So it looks as if there is a strong relationship between the variable indicating whether or not a student takes Statistics and the variable indicating whether or not a student takes Calculus. For example, a student who does not take Statistics is more likely to take Calculus than a student who does take Statistics. Of course, this is not a real data set, and it is small at that. So, in reality, we could expect a somewhat different conclusion.

Some graphical techniques are also available for this problem. In Figure 9.1.1, we have plotted the conditional distributions, given row, in a bar chart using the command Graph ▶ Bar Chart. This in turn leads to the dialog box shown in Display 9.1.2, where we have selected Cluster. This leads to the dialog box shown in Display 9.1.3, where we have entered the variables C2, C3 in the Categorical variables box (note order) and then clicked on Chart Options to bring up the dialog box shown in Display 9.1.4. Here we have indicated that we

want to display the distributions as percents. These plots are an evocative way to display the relationship between the variables.

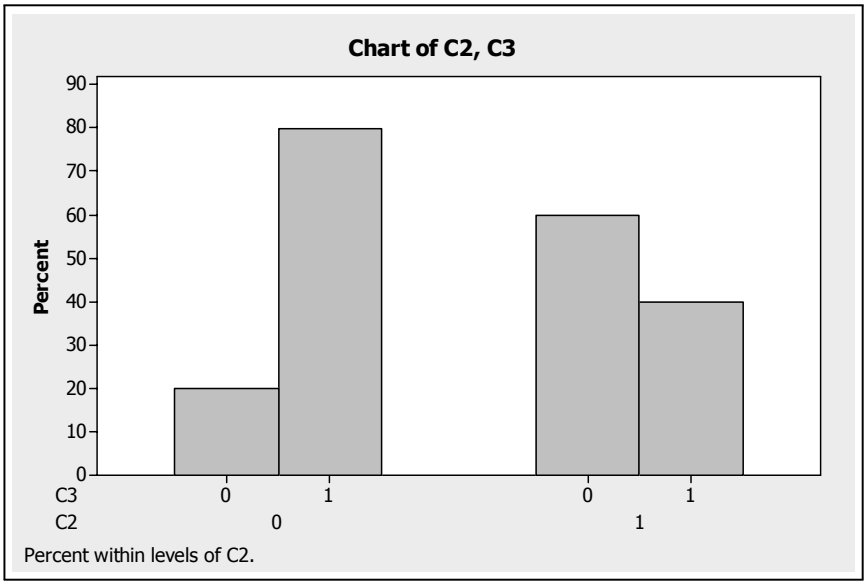
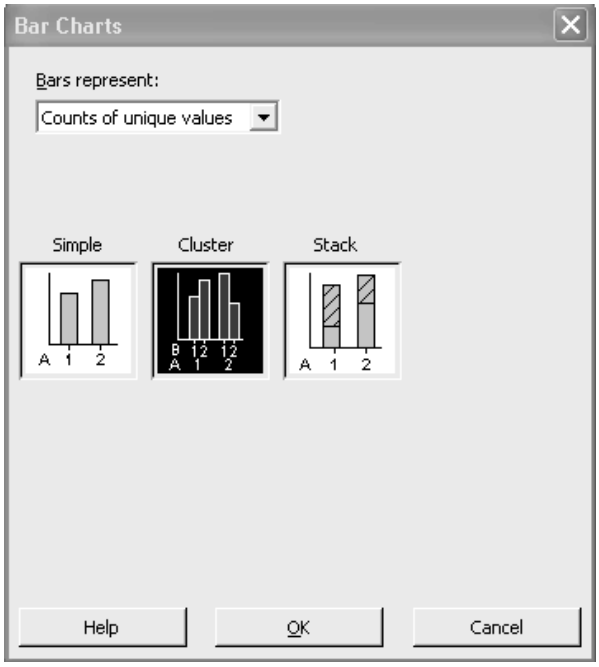
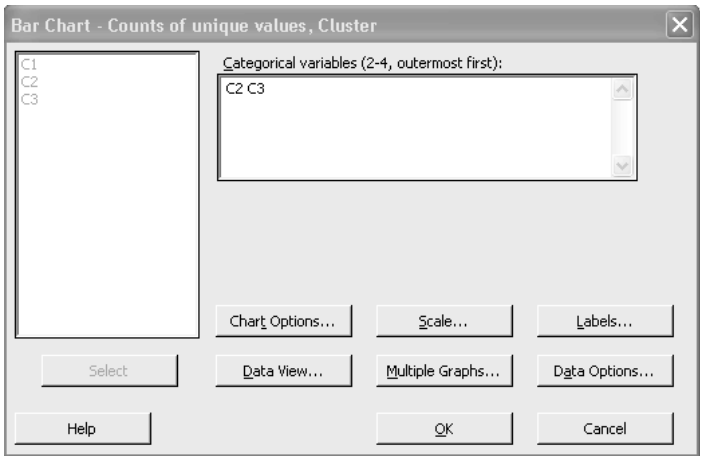


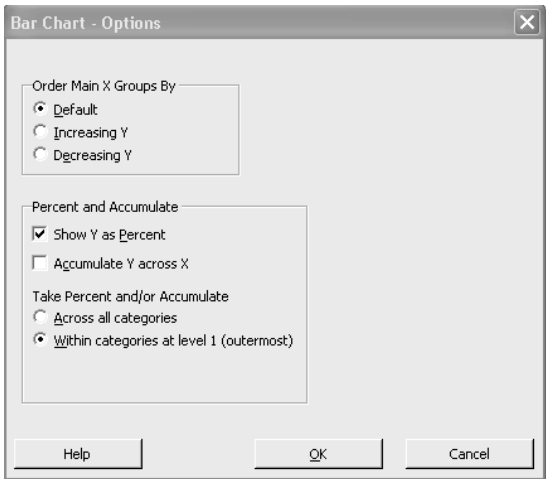
Figure 9.1.1: Conditional distributions of columns given row.



Display 9.1.2: Dialog box for selecting type of bar chart.



Display 9.1.3: Dialog box for choosing variables to graph in bar chart.



Display 9.1.4: Dialog box for selecting options for the bar chart.

The corresponding session command is **table** and there are the subcommands **totpercents, rowpercents,** and **colpercents** to specify whether or not we want total percents, row percents, and column percents to be printed for each cell. For example,

```
MTB > table c2 c3
```

produces the table of counts shown previously. If you do not want the marginal statistics to be printed, use the **noall** subcommand. Any cases with missing values are not included in the cross-tabulation. If you want them to be included, use the **missing** subcommand and a row or column will be printed, whichever is relevant, for missing values. For example, the subcommand

```
SUBC> missing c2 c3;
```

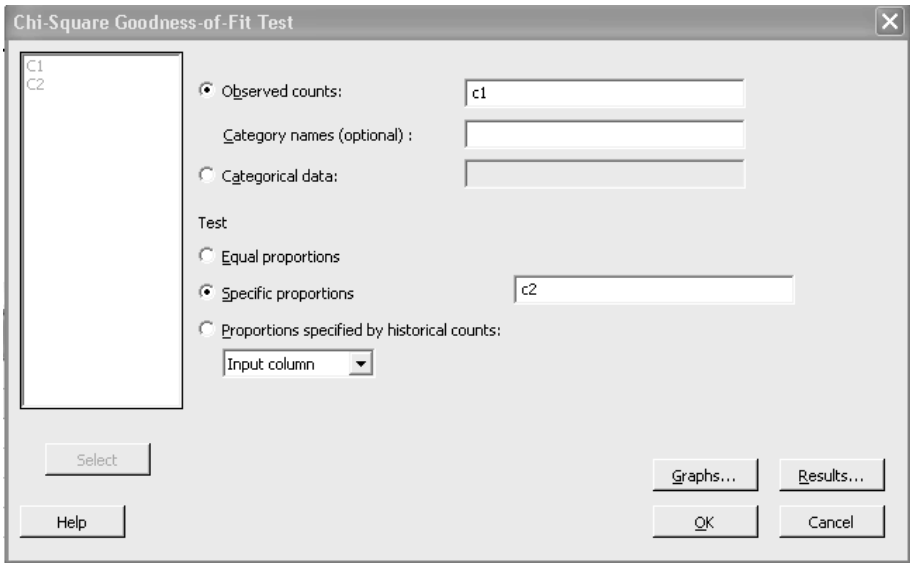
ensures that any cases with missing values in C2 or C3 are also tabulated.

9.2 The Chi-square Test

If you have a single variable, you use the Stat ▶ Tables ▶ Chi-Square Goodness-of-Fit Test (One variable) command to carry out a chi-square goodness-of-fit test to test whether or not you have a sample from specific distribution. For example, suppose we have a response that takes three possible values—namely, $1, 2$, and 3—and p_i is the probability that a response equals i. Suppose we want to test the hypothesis $H_0 : p_1 = .2, p_2 = .4, p_3 = .4$ and we have a sample of 100 where 22 values equal 1, 45 values equal 2, and 23 values equal 3. The counts are stored in C1 and hypothesized probabilities are stored in C2. Then the Stat ▶ Tables ▶ Chi-Square Goodness-of-Fit Test (One variable) command together with the dialog box shown in Display 9.2.1 gives the output

```
   Chi-Square Goodness-of-Fit Test for Observed Counts in Variable:
C1

                          Test            Contribution
   Category Observed Proportion Expected   to Chi-Sq
   1            22       0.2       20       0.200
   2            45       0.4       40       0.625
   3            33       0.4       40       1.225
    N  DF  Chi-Sq  P-Value
   100 2    2.05    0.359
```

and with the P-value equal to 0.359 we do not reject H_0. The output also includes a Chart of Observed and Expected Values and a Chart of Contribution to the Chi-Square Value by Category. These barcharts can be helpful in identifying where deviations from H_0 arise when we have evidence against H_0.

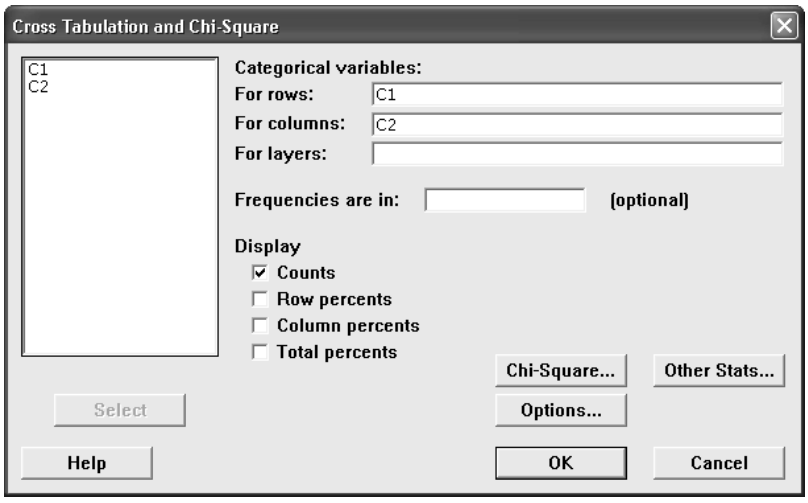


Display 9.2.1: Dialog box for a chi-square test on a single variable.

Now suppose we have more than one variable and we are interested in whether or not a relationship exists among these variables. Recall that there is no relationship between the variables—i.e., the variables are *independent*—if and only if the conditional distributions of one variable given the other are all the same. So in a two-way table we can assess whether or not there is a relationship by comparing the observed conditional distributions of the columns given the rows. Of course, there will be differences in these conditional distributions simply due to sampling error as we only have a sample and we are only estimating the true conditional distributions. Whether or not these differences are significant is assessed by conducting a chi-square test. When the table has r rows and c columns and we are testing for independence, then $k = (r-1)(c-1)$. Note that for a cell, the square of a *cell's standardized residual* is that cell's contribution to the chi-square statistic, namely

$$\frac{(\text{observed count in cell} - \text{expected count in cell})^2}{\text{expected count in cell}}$$

For example, suppose for 60 cases we have a categorical variable in C1 taking the values 0 and 1 and a categorical variable in C2 taking the values 0, 1 and 2. Consider the Stat ▶ Tables ▶ Cross Tabulation and Chi-Square command with the dialog box as in Display 9.2.2.



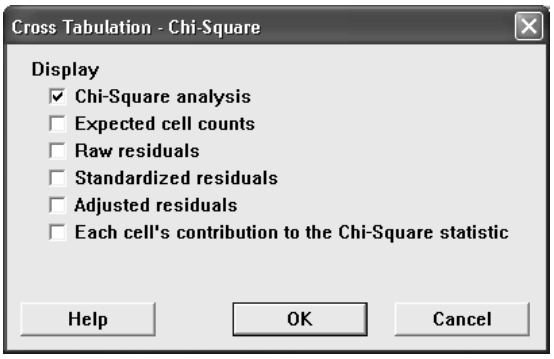
Display 9.2.2: Dialog box for cross-tabulating categorical variables.

Then clicking on the OK button produced the table

```
Rows:  C1     Columns:  C2
            0       1       2     All
0          10      13      11      34
1           9      10       7      26
All        19      23      18      60
Cell Contents:                      Count
```

in the Session window. This records the counts in the 6 cells of a table, with C1 indicating row and C2 indicating column. The variable C1 could be indicating a population with C2 a categorical variable defined on each population (or conversely), or both variables could be defined on a single population.

When using the Stat ▶ Tables ▶ Cross Tabulation and Chi-Square command, a chi-square analysis can be carried out by clicking the Chi-Square box in the dialog box of Display 9.2.2 as this brings up the dialog box shown in Display 9.2.3 where we have checked the Chi-square analysis box.



Display 9.2.3: Dialog box for carrying out a chi-square analysis.

The remaining boxes give additional options concerning what is printed in the Session window. We have chosen to have only the cell count printed for each cell in addition to the chi-square statistic and its associated P-value. Clicking on the OK button in this dialog box leads to the output

```
Rows:  C1    Columns:  C2
           0       1       2      All
0         10      13      11       34
1          9      10       7       26
All       19      23      18       60

Cell Contents:                          Count
Pearson Chi-Square = 0.271, DF = 2, P-Value = 0.873
```

being printed in the Session window. The P-value for testing the null hypothesis that these two categorical variables are independent against the alternative that they are not independent is .873, and so we do not reject the null hypothesis.

It is possible to cross-tabulate more than two variables and to test simultaneously for mutual statistical independence among the variables using the Stat ▶ Tables ▶ Cross Tabulation and Chi-Square command. Recall that it is also a good idea to plot the conditional distributions as well.

The general syntax of the corresponding session command **table** command is

table $\mathrm{E}_1 \ldots \mathrm{E}_m$;
chisquare V.

where E_1, ..., E_m are columns containing categorical variables and V is either omitted or takes the value 1, 2, or 3. The value 1 is the default and causes the count to be printed in each cell and can be omitted. The value 2 causes the count and the expected count, under the hypothesis of independence, to be printed in each cell. The value 3 causes the count, the expected count, and the standardized residual to be printed in each cell. For example, the command

```
MTB > table c1 c2;
SUBC> chisquare.
```

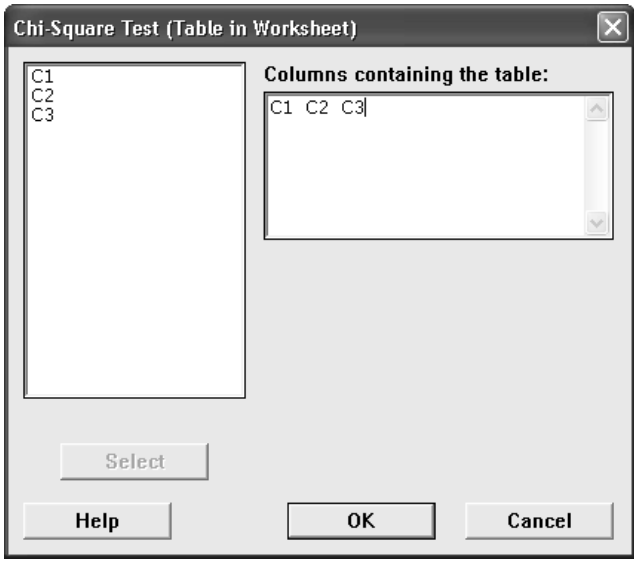
also produces the above output.

9.3 Analyzing Tables of Counts

If you have a two-way cross-tabulation for which the cell counts are already tabulated, you can use the Stat ▶ Tables ▶ Chi-Square Test (Two-Way Table in Worksheet) command on this data to carry out the chi-square analysis. For example, suppose we put the following data in columns C1–C3 as

Row	C1	C2	C3
1	51	22	43
2	92	21	28
3	68	9	22

corresponding to the counts arising from the cross-classification of a row and column variable. We then use the command Stat ▶ Tables ▶ Chi-Square Test (Two-Way Table in Worksheet) on this data with the dialog box as shown in Display 9.3.1.



Display 9.3.1: Dialog box for chi-square test on a table of counts.

This produces the output

```
Expected counts are printed below observed counts
         C1     C2     C3     Total
   1     51     22     43       116
      68.75  16.94  30.30
      4.584  1.509  5.320
   2     92     21     28       141
      83.57  20.60  36.83
      0.850  0.008  2.119
   3     68      9     22        99
      58.68  14.46  25.86
      1.481  2.062  0.577
Total   211     52    93       356
Chi-Sq = 18.510, DF = 4, P-Value = 0.001
```

in the Session window. The chi-square statistic has the value 18.51 in this case and the P-value is .001, so we reject the null hypothesis that there is no relationship between the row and column variables.

The general syntax of the corresponding session command **chisquare** command is

chisquare $\mathrm{E}_1 \ldots \mathrm{E}_m$

and this computes the expected cell counts, the chi-square statistic, and the associated P-value for the table in columns E_1, ..., E_m. Note that there is a limitation on the number of columns; namely we must have $m \leq 7$. For example, the command

```
MTB > chisquare c1-c3
```

produces the above chi-square analysis.

9.4 Exercises

1. Suppose that the observations in the following table are made on two categorical variables where variable 1 takes 2 values and variable 2 takes 3 values. Using the Stat ▶ Tables ▶ Cross Tabulation and Chi-Square command, cross-tabulate this data in a table of frequencies and in a table of relative frequencies. Calculate the conditional distributions of variable 1, given variable 2. Plot the conditional distributions. Is there any indication of a relationship existing between the variables? How many conditional distributions of variable 2, given variable 1, are there?

Obs	1	2	3	4	5	6	7	8	9	10
Var 1	0	0	0	1	1	0	1	0	0	1
Var 2	2	1	0	0	2	1	2	0	1	1

2. Suppose that two variables X and Y are cross-tabulated giving rise to the following table.

	$Y = 1$	$Y = 2$
$X = 1$	3388	389
$X = 2$	5238	1164
$X = 3$	1703	1699
$X = 4$	762	2045

   Use Minitab commands to calculate the marginal distributions of X and Y, and the conditional distributions of Y given X. Note that you cannot use _Stat ▶ Tables ▶ Cross Tabulation and Chi-square for this. Plot the conditional distributions.

3. Use Minitab to directly compute the expected frequencies, standardized residuals, chi-square statistic, and P-value for the hypothesis of independence in the table of Exercise 9.2.

4. Generate 1000 values from a uniform distribution on the integers $0, 1, \ldots, 9$. Use the Stat ▶ Tables ▶ Chi-Square Goodness-of-Fit Test (One variable) command to test that this generator is working correctly.

5. Suppose that we have the following table.

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	256	38	121
$X = 2$	47	54	73
$X = 3$	203	111	23

   Calculate and compare the conditional distributions of Y given X. Plot these conditional distributions in bar charts. Carry out a chi-square analysis to determine whether or not the variables in this problem are related.

6. Suppose we have a discrete distribution on the integers $1, \ldots, k$ with probabilities $p_1, \ldots, p_k$. Further, suppose we take a sample of n from this distribution and record the counts $f_1, \ldots, f_k$, where f_i records the number of times we observed i. It can be shown that

$$P(f_1 = n_1, \ldots, f_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

   when the n_i are nonnegative integers that sum to n. This is called the Multinomial$(n, p_1, \ldots, p_k)$ distribution, and it is a generalization of the Binomial(n, p) distribution. It is the relevant distribution for describing the counts in cross-tabulations. For $k = 4, p_1 = p_2 = p_3 = p_4 = .25, n = 3$, calculate these probabilities and verify that it is a probability distribution. Note that the gamma function is available with the Calc ▶ Calculator command (see Appendix B.1), and this can be used to evaluate factorials such as $n!$ and also $0! = 1$.

7. Calculate $P(f_1 = 3, f_2 = 5, f_3 = 2)$ for the Multinomial$(10, .2, .5, .3)$ distribution.

8. Generate (f_1, f_2, f_3) from the Multinomial$(1000, .2, .4, .4)$ distribution. Hint: Generate a sample of 1000 from the discrete distribution on 1, 2, 3 with probabilities .2, .4 , .4, respectively.

Chapter 10

Inference for Regression

New Minitab command discussed in this chapter

Stat ▶ Regression ▶ Residual Plots

This chapter deals with inference for the simple regression model. A regression analysis can be carried out using the command Stat ▶ Regression ▶ Regression. The regression as well as a scatterplot with the least-squares line overlaid can be obtained via Stat ▶ Regression ▶ Fitted Line Plot. Some aspects of these commands were discussed in II.2.3. Residual plots can be obtained using Stat ▶ Regression ▶ Residual Plots, provided you have saved the residuals.

10.1 Simple Regression Analysis

The command Stat ▶ Regression ▶ Regression provides a fit of the model $y = \alpha + \beta x + \epsilon$. Here, y is the *response variable,* x is the *explanatory* or *predictor variable,* ϵ is the *error variable* with an $N(0, \sigma)$ distribution, and α, β, and, σ are fixed unknown constants. These assumptions imply that, given x, the distribution of y is distributed $N(\alpha + \beta x, \sigma)$. So the mean of y given x is $\alpha + \beta x$, and this gives the relationship between y and x; i.e., as x changes at most the mean of the conditional distribution of y given x changes according to the linear function $\alpha + \beta x$.

The primary aim of a regression analysis is to make inferences about the unknown intercept α and the unknown slope β and to make predictive inferences about future values of y at possibly new values of x. All inferences are dependent on this model being correct. If we go ahead and report inferences when the model is incorrect, we run the risk of these inferences being invalid. So we must always check that the model makes sense in light of the data obtained. This is referred to as *model checking.*

We let $(x_1, y_1), \ldots, (x_n, y_n)$ denote the data on which we will base all our inferences. The basic inference method for this model is to use least-squares to estimate α and β, and we denote these estimates by a and b, respectively; i.e., a and b are the values of α and β that minimize

$$S^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 .$$

We predict the value of a future y, when the explanatory variable takes the value x, by $\hat{y} = a + bx$. The *ith fitted value* $\hat{y}_i$ is the estimate of the mean of y at x_i; i.e., $\hat{y}_i = a + bx_i$. The *ith residual* is given by $r_i = y_i - \hat{y}_i$; i.e., it is the error incurred when predicting the value of y at x_i by $\hat{y}_i$. We estimate the standard deviation σ by

$$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} r_i^2}$$

which equals the square root of the MSE (mean-squared error) for the regression model.

Of course, the estimates a, b, and $\hat{y}$ are not equal to the quantities that they are estimating. It is an important aspect of a statistical analysis to say something about how accurate these estimates are, and for this we use the standard error of the estimate. The standard error of a is given by

$$s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

The standard error of b is given by

$$s\sqrt{\frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

The standard error of the estimate $a + bx$ of the mean $\alpha + \beta x$ is given by

$$s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

To predict y at x, we must take into account the additional variation caused by the error ϵ, and so the standard error of $a + bx$, as a predictor of y at x, is given by

$$s\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Finally, the residual r_i, as an estimate of the error incurred at x_i, has standard error

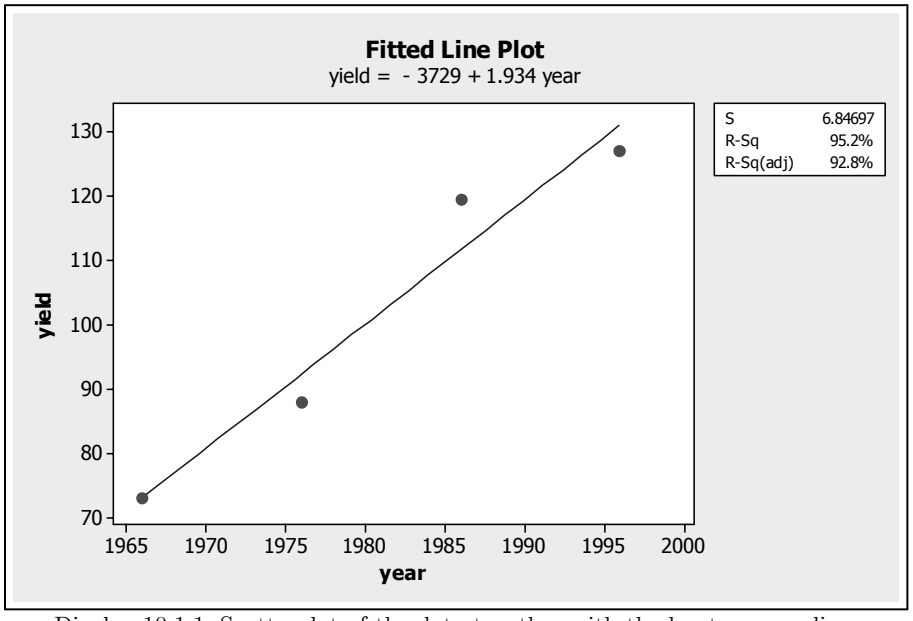
$$s\sqrt{1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}.$$

The *ith standardized residual* is then given by r_i divided by this quantity.

We now illustrate regression analysis using Minitab. Suppose we have the following data points

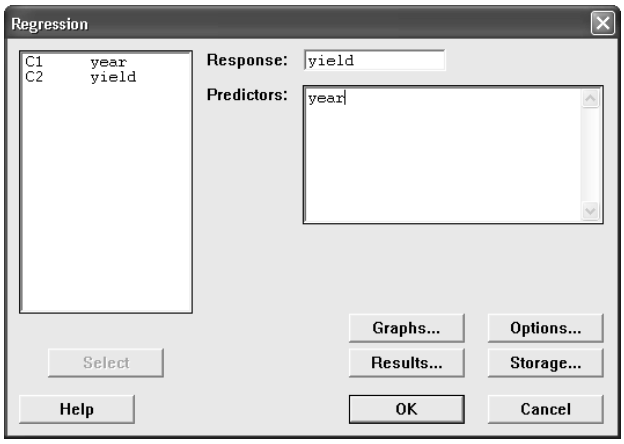
$$\begin{aligned}(x_1, y_1) &= (1966, 73.1)\\ (x_2, y_2) &= (1976, 88.0)\\ (x_3, y_3) &= (1986, 119.4)\\ (x_4, y_4) &= (1996, 127.1)\end{aligned}$$

where x is year and y is yield in bushels per acre and that we give x the name `year` and y the name `yield`. The Stat ▶ Regression ▶ Fitted Line Plot command with `yield` as the response and `year` as predictor produces the plot of Display 10.1.1

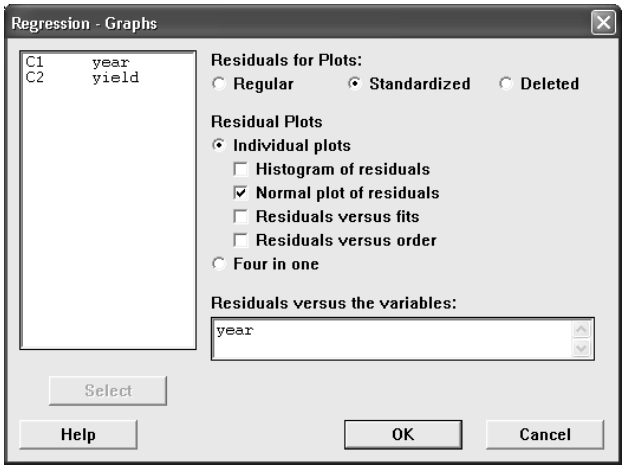


Display 10.1.1: Scatterplot of the data together with the least-squares line.

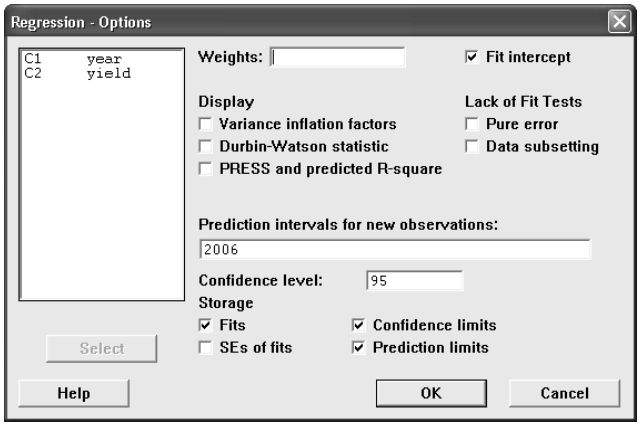
The Stat ▶ Regression ▶ Fitted Line Plot command also produces some of the Session window output below in the Session window. Because we wanted more features of a regression analysis than this command provides, we resorted to the Stat ▶ Regression ▶ Regression command together with the dialog boxes as in Displays 10.1.2, 10.1.3, 10.1.4, and 10.1.5.



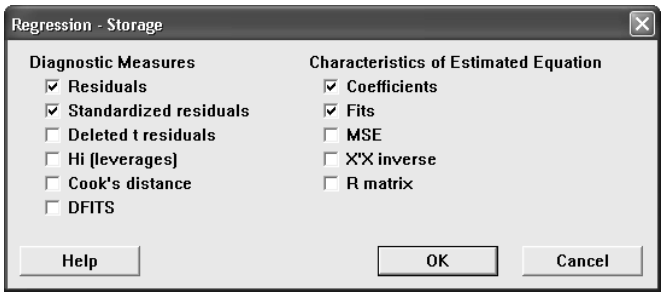
Display 10.1.2: Dialog box for simple regression analysis.



Display 10.1.3: Dialog box for selecting graphs to be plotted in regression analysis.



Display 10.1.4: Dialog box for selecting predictive inferences in a regression analysis.



Display 10.1.5: Dialog box for selecting quantities to be stored in a regression analysis.

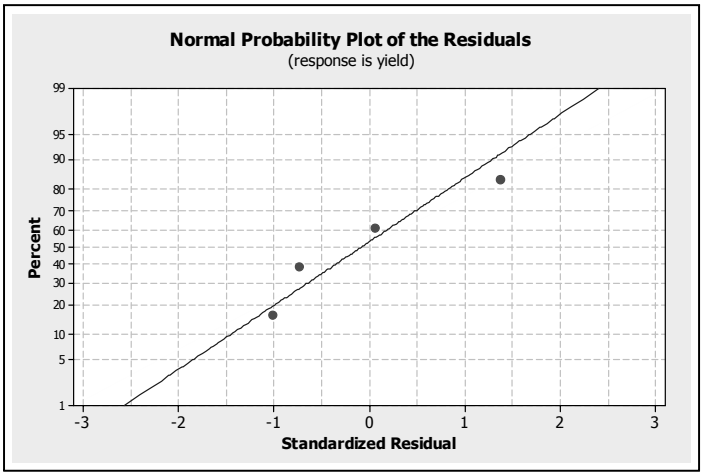
These entries in the dialog boxes produce the output

```
Regression Analysis:  yield versus year
The regression equation is
yield = - 3729 + 1.93 year

Predictor      Coef    SE Coef        T        P
Constant    -3729.4    606.6       -6.15    0.025
year         1.9340   0.3062        6.32    0.024

S = 6.84697 R-Sq = 95.2% R-Sq(adj) = 92.8%

Analysis of Variance
Source            DF         SS       MS        F        P
Regression         1     1870.2   1870.2    39.89    0.024
Residual Error     2       93.8     46.9
Total              3     1963.9

Predicted Values for New Observations
New
Obs    Fit   StDev Fit       95.0% CI          95.0% PI
 1  150.25        8.39  (114.17, 186.33)   (103.67, 196.83) X
X denotes a row with X values away from the center
```

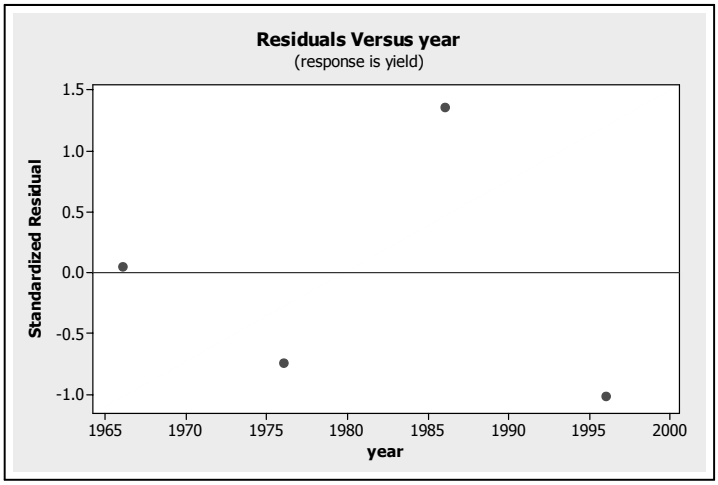
in the Session window.

The dialog box of Display 10.1.2 establishes that `yield` is the response and `year` is the explanatory variable. The output from this gives the least-squares line as $y = -3729 + 1.93x$. Further, the standard error of $a = -3729.4$ is 606.6, the standard error of $b = 1.934$ is 0.3062, the t statistic for testing $H_0 : \alpha = 0$ versus $H_a : \alpha \neq 0$ is -6.15 with P-value 0.025, and the t statistic for testing $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ is 6.32 with P-value 0.024. The estimate of σ is $s = 6.847$ and the squared correlation—coefficient of determination—is $R^2 = .952$, indicating that 95.2% of the observed variation in y is explained by the changes in x. The Analysis of Variance table indicates that the F statistic for testing $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ is 39.89 with P-value 0.024 and the MSE is 46.9. So we definitely reject the null hypothesis that there is no relationship between the response and the predictor.

Before clicking on the OK button of the dialog box of Display 10.1.2, however, we clicked on the Graphs button to bring up the dialog box of Display 10.1.3, the Options button to bring up the dialog box of Display 10.1.4, and the Storage button to bring up the dialog box of Display 10.1.5. In the Graphs dialog box, we specified that we want the standardized residuals plotted in a normal probability plot and plotted against the variable `year`. These plots appear in Displays 10.1.6 and 10.1.7, respectively, and don't indicate that any model assumptions are being violated.



Display 10.1.6: Normal probability plot of the standardized residuals.



Display 10.1.7: A plot of the standardized residuals versus the explanatory variable.

In the Options dialog box, we specified that we wanted to estimate the mean value of y at $x = 2006$ and report and store this value together with a 95% confidence interval for this quantity and a 95% prediction interval for a new observation at $x = 2006$. The output above gives the estimated mean value

at $x = 2006$ as 150.25 with standard error 8.39, and the 95% confidence and prediction intervals for this quantity are (114.17, 186.33) and (103.67, 196.83), respectively. The estimate is stored in the worksheet in a variable called `PFIT1`, and the endpoints of the confidence and prediction intervals are stored in the worksheet with the names `CLIM1`, `CLIM2`, `PLIM1`, `PLIM2`, respectively. In the Storage dialog box, we specified that we wanted to store the values of a and b, the fitted values, the residuals, and the standardized residuals. The residuals are stored in a variable called `RESI1`, the standardized residuals are stored in a variable called `SRES1`, the values of a and b are stored consecutively in a variable named `COEF1`, and the fitted values are stored in a variable called `FITS1`.

All of the stored quantities are available for further use. Suppose we want a 95% confidence interval for b. The commands

```
MTB > invcdf .975;
SUBC> student 2.
Student's t distribution with 2 DF
 P( X <= x)       x
    0.9750     4.3027
MTB > let k2=4.3027*.3062
MTB > let k3=coef1(2)-k2
MTB > let k4=coef1(2)+k2
MTB > print k3 k4
K3 0.616513
K4 3.25149
```

give this interval as $(0.617, 3.251)$.

The general syntax of the corresponding session command **regress** command for fitting a line is

regress E_1 1 E_2

where E_1 contains the values of the response variable y and E_2 contains the values of the explanatory variable x. There are a number of subcommands that can be used with **regress,** and these are listed and explained below.

coefficients E_1 — stores the estimates of the coefficients in column E_1.

constant (noconstant) — ensures that β_0 is included in the regression equation, while **noconstant** fits the equation without β_0.

fits E_1 — stores the *fitted values* $\hat{y}$ in E_1.

ghistogram — causes a histogram of the residuals specified in **rtype** to be plotted.

gfits — causes a plot of the residuals specified in **rtype** versus the fitted values to be plotted.

gnormal — causes a normal quantile plot of the residuals specified in **rtype** to be plotted.

gorder — causes a plot of the residuals specified in **rtype** versus order to be plotted.

gvariable E_1 — causes a plot of the residuals specified in **rtype** versus the explanatory variable in column E_1 to be plotted.

mse E_1 — stores the mean squared error in constant E_1.

predict $E_1 \ldots E_k$ — (k is the number of explanatory variables where $k = 1$ with simple linear regression) computes and prints the predicted values at E_1, ..., E_k, where these are columns of the same length or constants with E_i corresponding to the *ith* explanatory variable. Also, this prints the estimated standard deviations of these values, confidence intervals for these values, and prediction intervals. The subcommand **predict** in turn has a number of subcommands.

> **confidence** V — V specifies the level for the confidence intervals.
>
> **pfits** E_1 — stores the predicted values in E_1.
>
> **psdfits** E_1 — stores the estimated standard deviations of the predicted values in E_1.
>
> **climits** E_1 E_2 — stores the lower- and upper-confidence limits for the predicted values in E_1 and E_2, respectively.
>
> **plimits** E_1 E_2 — stores the lower- and upper-prediction limits for the predicted values in E_1 and E_2, respectively.

residuals E_1 — stores the regular residuals in E_1.

rtype V — indicates what type of residuals are to be used in the plotting subcommands, where V = 1 is the default and specifies regular residuals, V = 2 specifies standardized residuals, and V = 3 specifies Studentized deleted residuals.

sresiduals E_1 — stores the standardized residuals—the residuals divided by their estimated standard deviations—in E_1.

For example, the session commands

```
MTB > regress 'yield' 1 'year';
SUBC> coefficients c3;
SUBC> mse k1;
SUBC> fits c4;
SUBC> residuals c5;
SUBC> sresiduals c6;
SUBC> rtype 2;
SUBC> gnormal;
SUBC> gvariable 'year';
SUBC> predict 2006;
SUBC> pfits c7;
SUBC> climits c8 c9;
SUBC> plimits c10 c11.
```

produce the same results as the menu commands with the dialog boxes as in Displays 10.1.2, 10.1.3, 10.1.4, and 10.1.5 for the example of this section although the ordering of the columns is different.

10.2 Exercises

1. In C1, place the x values -3.0, -2.5, $-2.0, \ldots, 2.5, 3.0$. In C2, store a sample of 13 from the error ϵ, where ϵ is distributed $N(0, 2)$. In C3, store the values $y = \beta_0 + \beta_1 x + \epsilon = 1 + 3x + \epsilon$. Calculate the least-squares estimates of β_0 and β_1 and the estimate of σ^2. Repeat this example but take five observations at each value of x. Compare the estimates from the two situations and their estimated standard deviations.

2. In C1, place the x values -3.0, -2.5, $-2.0, \ldots, 2.5, 3.0$. In C2, store a sample of 13 from the error ϵ, where ϵ is distributed $N(0, 2)$. In C3, store the values $y = \beta_0 + \beta_1 x + \epsilon = 1 + 3x + \epsilon$. Plot the least-squares line. Repeat your computations twice after changing the first y observation to 20 and then to 50, and make sure the scales on all the plots are the same. What effect do you notice?

3. In C1, place the x values -3.0, -2.5, $-2.0, \ldots, 2.5, 3.0$. In C2, store a sample of 13 from the error ϵ, where ϵ is distributed $N(0, 2)$. In C3, store the values $y = \beta_0 + \beta_1 x + \epsilon = 1 + 3x + \epsilon$. Plot the standardized residuals in a normal quantile plot against the fitted values and against the explanatory variable. Repeat this, but in C3 place the values of $y = 1 + 3x - 5x^2 + \epsilon$. Compare the residual plots.

4. In C1, place the x values -3.0, -2.5, $-2.0, \ldots, 2.5, 3.0$. In C2, store a sample of 13 from the error ϵ, where ϵ is distributed $N(0, 2)$. In C3, store the values $y = \beta_0 + \beta_1 x + \epsilon = 1 + 3x + \epsilon$. Plot the standardized residuals in a normal quantile plot against the fitted values and against the explanatory variable. Repeat this, but in C2 place the values of a sample of 13 from the Student(1) distribution. Compare the residual plots.

5. In C1, place the x values -3.0, -2.5, $-2.0, \ldots, 2.5, 3.0$. In C2, store a sample of 13 from the error ϵ, where ϵ is distributed $N(0, 2)$. In C3, store the values $y = \beta_0 + \beta_1 x + \epsilon = 1 + 3x + \epsilon$. Calculate the predicted values and the lengths of .95 confidence and prediction intervals for this quantity at $x = .1, 1.1, 2.1, 3.5, 5, 10$, and 20. Explain the effect that you observe.

6. In C1, place the x values -3.0, -2.5, $-2.0, \ldots, 2.5, 3.0$. In C2, store a sample of 13 from the error ϵ, where ϵ is distributed $N(0, 2)$. In C3, store the values $y = \beta_0 + \beta_1 x + \epsilon = 1 + 3x + \epsilon$. Calculate the least-squares estimates and their estimated standard deviations. Repeat this, but for C1 the x values are to be 12 values of -3 and one value of 3. Compare your results and explain them.

Chapter 11

Multiple Regression

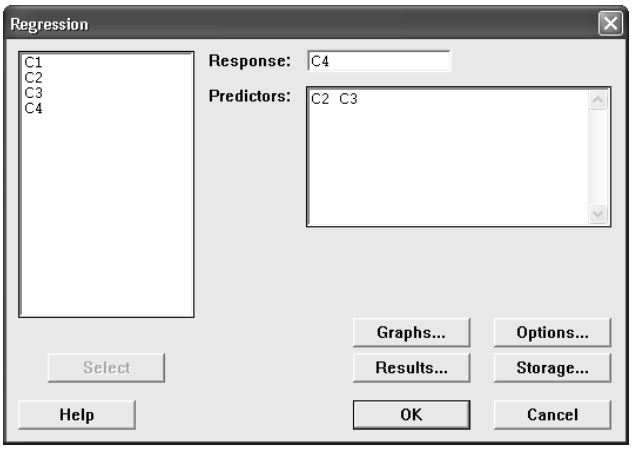
In this chapter, we discuss *multiple regression;* i.e., we have a single numeric response variable y and $k > 1$ explanatory variables $x_1, \ldots, x_k$. There are no real changes in the behavior of the Stat ▶ Regression ▶ Regression command, and the descriptions we gave in Chapter 10 apply to this chapter as well. We present an example of a multiple regression analysis using Minitab.

A multiple regression analysis can be carried out using Stat ▶ Regression ▶ Regression and filling in the dialog box appropriately. Residual plots can be obtained using Stat ▶ Regression ▶ Residual Plots provided you have saved the residuals. Also available in Minitab are *stepwise regression* using Stat ▶ Regression ▶ Regression ▶ Stepwise and *best subsets regression* using Stat ▶ Regression ▶ Regression ▶ Best Subsets.

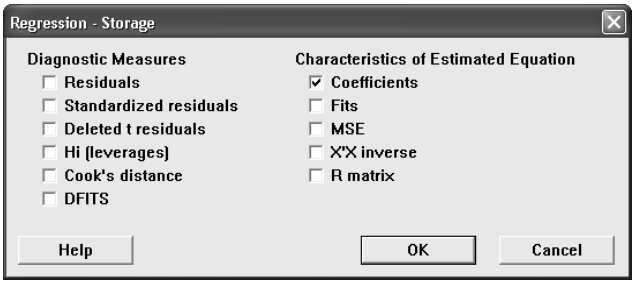
11.1 Example of a Multiple Regression

We consider a generated multiple regression example to illustrate the use of the Stat ▶ Regression ▶ Regression command in this context. Suppose that $k = 2$ and $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon = 1 + 2x_1 + 3x_2 + \epsilon$, where ϵ is distributed $N(0, \sigma)$ with $\sigma = 1.5$. We generated a sample of 16 from the $N(0, 1.5)$ distribution and placed these values in C1. In C2 we stored the values of x_1 and in C3 stored the values of x_2. Suppose that these variables take every possible combination of $x_1 = -1, -.5, .5, 1$ and $x_2 = -2, -1, 1, 2$. In C4, we placed the values of the response variable y.

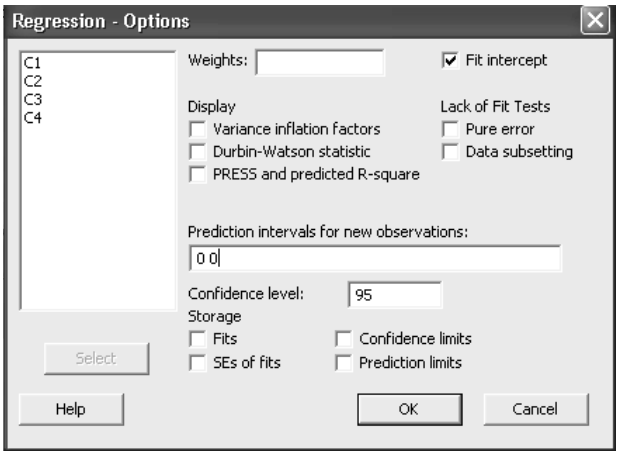
We then proceeded to analyze this data as if we didn't know the values of β_0, β_1, β_2, and σ. The Stat ▶ Regression ▶ Regression command as implemented in Display 11.1.1, together with Displays 11.1.2, 11.1.3 and 11.1.4,



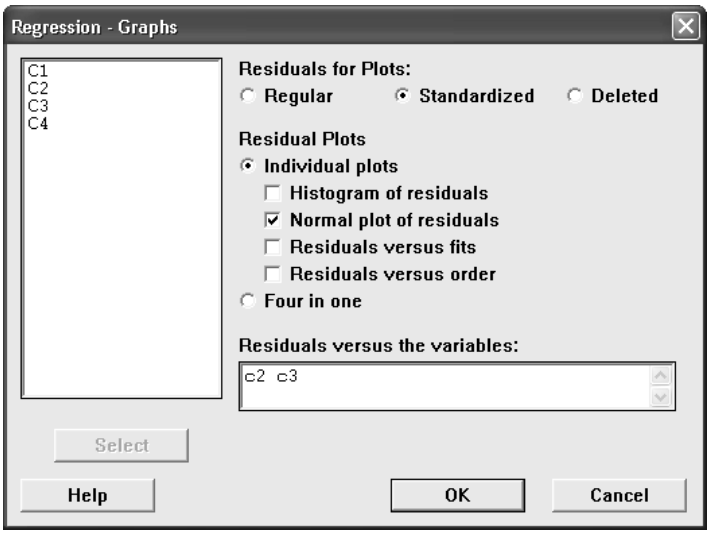
Display 11.1.1: Dialog box for the Stat ▶Regression ▶Regression command in the example.



Display 11.1.2: Dialog box obtained by clicking on the Storage button in the dialog box depicted in Display 11.1.1. We have requested that the least-squares coefficients be stored.



Display 11.1.3: Dialog box obtained by clicking on the Options button in the dialog box depicted in Display 11.2.1. We have requested that a value be predicted at the settings $x_1 = 0, x_2 = 0$.



Display 11.1.4: Dialog box obtained by clicking on the Graphs button in the dialog box depicted in Display 11.1.1.

produces the output in the Session window

```
Regression Analysis:  C4 versus C2, C3
The regression equation is
C4 = 0.996 + 2.32 C2 + 3.25 C3
Predictor      Coef    SE Coef       T         P
Constant     0.9958     0.2811    3.54     0.004
C2           2.3162     0.3556    6.51     0.000
C3           3.2517     0.1778   18.29     0.000
S = 1.12451    R-Sq = 96.7%    R-Sq(adj) = 96.2%
Analysis of Variance
Source          DF        SS        MS       F       P
Regression       2    476.58    238.29  188.44   0.000
Residual Error  13     16.44      1.26
Total           15    493.02
Source    DF    Seq SS
C2         1     53.65
C3         1    422.93
Predicted Values for New Observations
New
Obs    Fit    SE Fit        95% CI             95% PI
 1   0.996     0.281    (0.388, 1.603)    (-1.508, 3.500)
Values of Predictors for New Observations
New
Obs              C2                C3
 1         0.000000          0.000000
```

and the plot in Display 11.1.5. This specifies the least-squares equation as $y = 0.996 + 2.32x_1 + 3.25x_2$. For example, the estimate of β_1 is $b_1 = 2.3162$ with standard error 0.3556 and the t statistic for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is 6.51 with P-value 0.000. The estimate of σ is $s = 1.12451$ and $R^2 = .967$. The Analysis of Variance table indicates that the F statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ takes the value 188.44 with P-value 0.000 so we would definitely reject the null hypothesis. Also, the MSE is given as 1.26.

The table after the Analysis of Variance table is called the *Sequential Analysis of Variance table* and is used when we want to test whether or not explanatory variables are in the model in a prescribed order. For example, the table that contains the rows labeled C2 and C3 allows for the testing of the sequence of hypotheses $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$ and—if we reject this (and only if we do)—then testing the hypothesis $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. To test these hypotheses, we first compute $F = 422.93/s^2 = 422.93/1.26 = 335.66$ and then compute the P-value $P(F(1, 13) > 335.66) = 0.000$, and so we reject and go no further. If we had not rejected this null hypothesis, the second null hypothesis would be tested in exactly the same way using $F = 53.65/1.26 = 42.58$. Obviously, the order in which we put variables into the model matters with these sequential tests. Sometimes, it is clear how to do this; for example, in fitting a quadratic model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ we put $x_1 = x$ and $x_2 = x^2$ and test for the existence of the quadratic term first and, if no quadratic term is found, test for the existence of the linear term. Sometimes, the order for testing is not as clear and the sequential tests are not as appropriate.

The dialog box in Display 11.1.2 is obtained by clicking on the Storage button in the dialog box of Display 11.1.1. We stored the values of the least-squares estimates in C5, as the dialog box in Display 11.1.2 indicates, and so these are available for forming confidence intervals. Then, for example, the commands

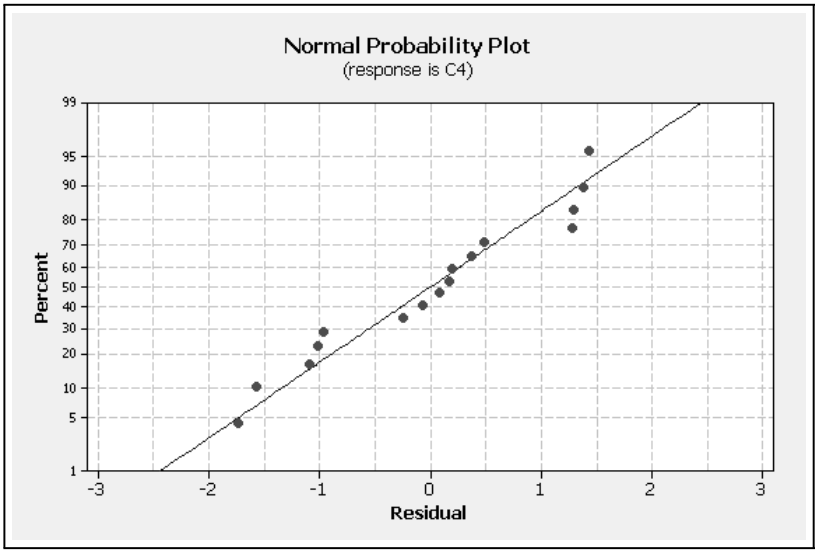
```
MTB > invcdf .95;
SUBC> student 13.
Student's t distribution with 13 DF
 P(X <= x) x
 0.9500 1.7709
MTB > let k1=1.7709*0.1778
MTB > let k2=c5(3)-k1
MTB > let k3=c5(3)+k1
MTB > print k2 k3
K2 2.93681
K3 3.56654
```

compute a 90% confidence interval for β_2 as $(2.93681, 3.56654)$, which we note does cover the true value in this case.

The dialog box in Display 11.1.3 is obtained by clicking on the Options button in the dialog box of Display 11.1.1. The dialog box in Display 11.1.3 indicates that we requested that the program compute the predicted value at $x_1 = 0, x_2 = 0$ as well as the confidence and prediction intervals for this value.

We obtained the predicted value as 0.996 with standard error 0.281 and as well the 95% confidence and prediction intervals given by $(0.388, 1.603)$ and $(-1.508, 3.500)$, respectively.

The dialog box in Display 11.1.4 is obtained by clicking on the Graphs button in the dialog box of Display 11.1.1. Here we requested a normal quantile plot of the standardized residuals, which we show in Display 11.1.5, and also requested plots of the standardized residuals against each of the explanatory variables, which we don't show. All of these plots look reasonable.

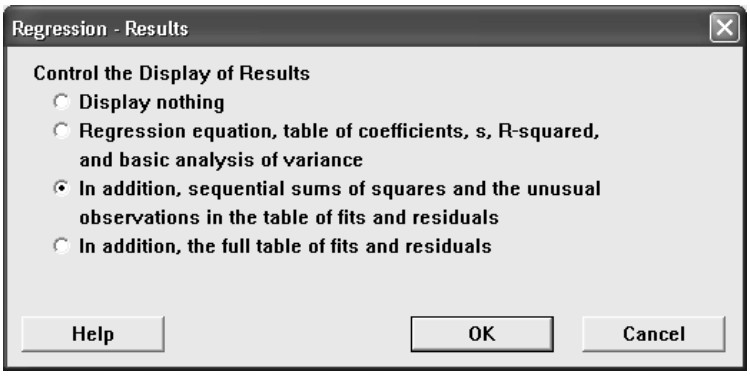


Display 11.1.5: Normal probability plot of the standardized residuals for the example.

The following session commands produce the above output for the example of this section.

```
MTB > regress c4 2 c2 c3;
SUBC> coefficients c5;
SUBC> rtype 2;
SUBC> gnormal;
SUBC> gvariable C2 C3;
SUBC> predict 0 0;
SUBC> climits c6 c7;
SUBC> plimits c8 c9.
```

We can also control the amount of output obtained from the Stat ▶ Regression ▶ Regression command. This is accomplished by clicking on the Results button of the dialog box shown in Display 11.1.1 bringing up Display 11.1.6. We have requested that, in addition to the fitted regression equation, least-squares coefficients, s, R^2, and ANOVA table, the table of sequential sums of squares (for the order in which the variables appear in the model), and a table of unusual observations be printed.



Display 11.1.6: Dialog box obtained by clicking on the Results button in the dialog box depicted in Display 11.1.1.

The session command to control the amount output from the **regress** and other Minitab commands is **brief.** The general syntax of the **brief** command is

brief V

where V is a nonnegative integer that controls the amount of output. For any given command the output is dependent on the specific command although V = 0 suppresses all output, for all commands, beyond error messages and warnings. The default level of V is 2. When V = 3, the **regress** command produces the usual output and in addition prints x, y, $\hat{y}$, the standard deviation of $\hat{y}$, $y - \hat{y}$ and the standardized residual. When V = 1, the regress command gives the same output as when V = 2 but the sequential analysis of variance table is not printed. Don't forget that after you set the level of **brief,** this may affect the output of all commands you subsequently type and therefore it may need to be reset.

11.2 Exercises

1. In C1, place the x_1 values $-3.0, -2.5, -2.0, \ldots, 2.5, 3.0$. In C2, store a sample of 13 from the error ϵ, where ϵ is distributed $N(0, 2)$. In C3, store the values of $x_2 = x^2$. In C4 store the values of $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon = 1 + 3x + 5x^2 + \epsilon$. Calculate the least-squares estimates of β_0, β_1, and β_2, and the estimate of σ^2. Carry out the sequential F tests testing first for the quadratic term and then, if necessary, testing for the linear term.

2. In C1, place the x values $-3.0, -2.5, -2.0, \ldots, 2.5, 3.0$. In C2, store a sample of 13 from the error ϵ, where ϵ is distributed $N(0, 2)$. Fit the model $y = 1 + 3\cos(x) + 5\sin(x) + \epsilon$. Calculate the least-squares estimates of β_0, β_1, and β_2, and the estimate of σ^2. Carry out the F test for any effect due to x. Are the sequential F tests meaningful here?

3. In C1, place the x_1 values $-3.0, -2.5, -2.0, \ldots, 2.5, 3.0$. In C2, store a sample of 13 from the error ϵ, where ϵ is distributed $N(0, 2)$. In C3, store the values of $x_2 = x^2$. In C4, store the values of $y = 1+3\cos(x)+5\sin(x)+\epsilon$. Next fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ and plot the standardized residuals in a normal quantile plot and against each of the explanatory variables.

Chapter 12

One-Way Analysis of Variance

New Minitab commands discussed in this chapter

Stat ▶ ANOVA ▶ One-way
Stat ▶ ANOVA ▶ One-way (Unstacked)

This chapter deals with methods for making inferences about the relationship existing between a single numeric response variable and a single categorical explanatory variable. The basic inference methods are the one-way analysis of variance (ANOVA) and the comparison of means. There are two commands for carrying out a one-way analysis of variance, namely Stat ▶ ANOVA ▶ One-way and Stat ▶ ANOVA ▶ One-way (Unstacked). They differ in the way the data must be stored for the analysis.

We write the one-way ANOVA model as $x_{ij} = \mu_i + \epsilon_{ij}$, where $i = 1, \ldots, I$ indexes the levels of the categorical explanatory variable and $j = 1, \ldots, n_i$ indexes the individual observations at each level, μ_i is the mean response at the *ith* level, and the errors ϵ_{ij} are a sample from the $N(0, \sigma)$ distribution. Based on the observed x_{ij}, we want to make inferences about the unknown values of the parameters $\mu_1, \ldots, \mu_I, \sigma$.

12.1 A Categorical Variable and a Quantitative Variable

Suppose that we have two variables—one is categorical and one is quantitative—and we want to examine the form of the relationship between these variables. Of course there may not even be a relationship between the variables. We treat the situation where the categorical variable is explanatory and the quantitative

variable is the response and examine some basic techniques for addressing this question.

To illustrate, we use the data in the following table. Here, we have four different colors of insect trap—lemon yellow, white, green, and blue—and the number of insects trapped on six different instances of each trap.

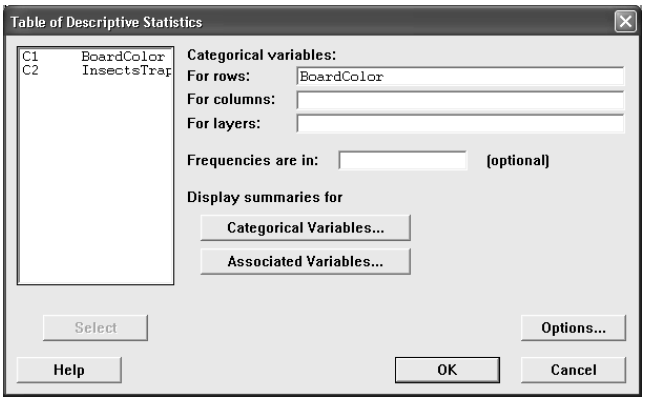
Board Color	Insects Trapped
Lemon Yellow	45 59 48 46 38 47
White	21 12 14 17 13 17
Green	37 32 15 25 39 41
Blue	16 11 20 21 14 7

We have read these data into a worksheet so that C1 contains the trap color, and labelled this column `BoardColor`, with 1 indicating lemon yellow, 2 indicating white, 3 indicating green, and 4 indicating blue, and in C2 we have put the numbers of insects trapped, and labelled this column `InsectsTrapped`. We calculate the mean number of insects trapped for each trap using the Stat ▶ Tables ▶ Descriptive Statistics command with the dialog boxes as in Displays 12.1.1 and 12.1.2. In the dialog box of Display 12.1.1, we have put `BoardColor` into the Categorical variables: For rows box and clicked on the Associated Variables button to bring up the dialog box of Display 12.1.2. In this box, we have put `InsectsTrapped` into the Associated variables box and selected Means to indicate that we want the mean of C2 to be computed for each value of C1. Clicking on the OK buttons produces the output

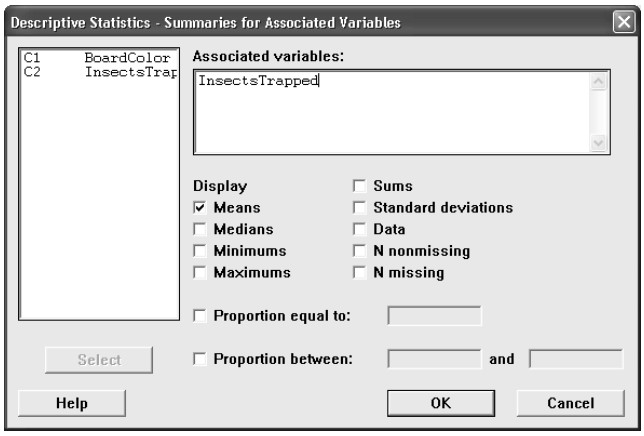
```
 InsectsTrapped
        Mean    Count
1    47.17          6
2    15.67          6
3    31.50          6
4    14.83          6
All  27.29         24
```

in the Session window. The fact that the means change from one level of C1 to another seems to indicate that there is some relationship between the color of insect trap and the number of insects trapped. As indicated in Display 12.1.1, there are many other statistics, besides the mean, that we could have chosen to tabulate.

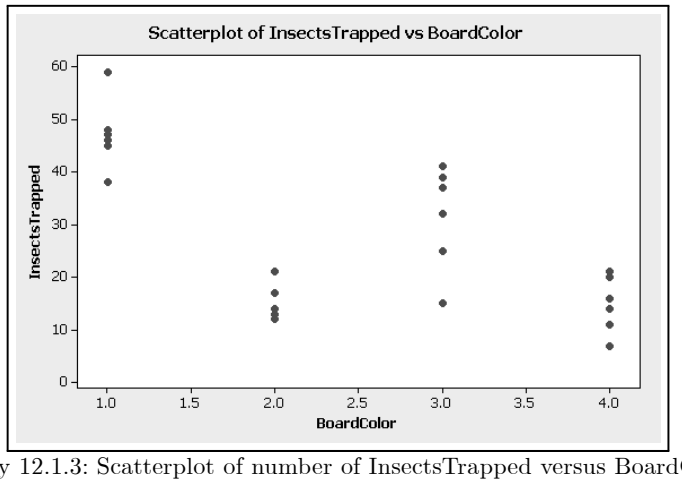
It is also a good idea to look at a scatterplot of the quantitative variable versus the categorical variable. We can do this with Graph ▶ Scatterplot and obtain the plot shown in Display 12.1.3.



Display 12.1.1: First dialog box for tabulating a quantitative variable by a categorical variable.

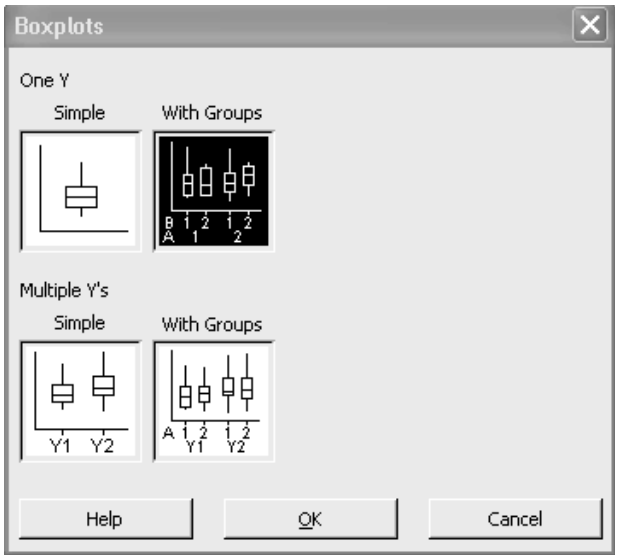


Display 12.1.2: Second dialog box for tabulating a quantitative variable by a categorical variable.

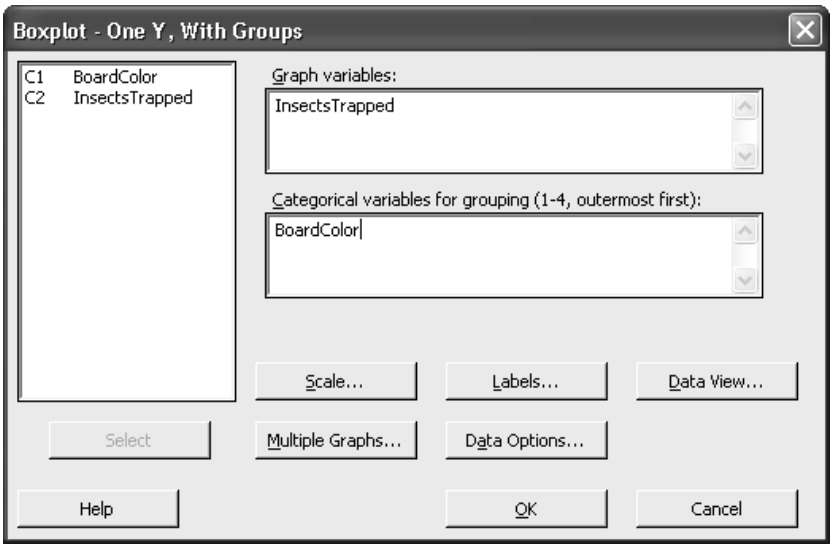


Display 12.1.3: Scatterplot of number of InsectsTrapped versus BoardColor.

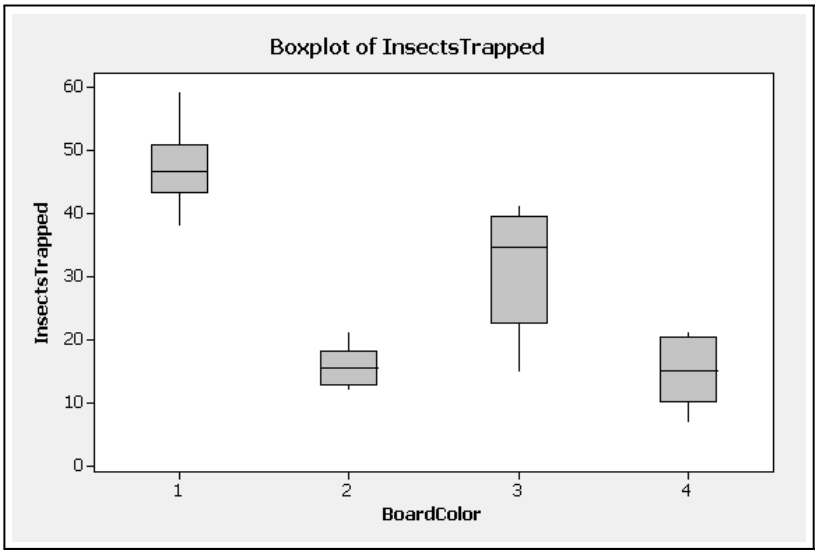
Another useful plot in this situation is to create side-by-side boxplots. This can be carried out using the Graph ▶ Boxplot command. The dialog box of Display 12.1.4 permits the choice of type of boxplot to plot and we have selected With Groups. In the dialog box of Display 12.1.5 we have put `InsectsTrapped` in the Graph Variables box and `BoardColor` in the Categorical variables for grouping box. Clicking on the OK buttons produces the plot shown in Display 12.1.6.



Display 12.1.4: Dialog box for selecting type of boxplot.



Display 12.1.5: Dialog box for creating side-by-side boxplots.



Display 12.1.6: Side-by-side boxplots.

The session command **table** can also be used for creating the tables we have described in this section. For the example as described above, the commands

```
MTB > table c1;
SUBC> means c2.
```

produce the mean number of insects trapped for each color of trap as given above. Besides the **means** subcommand, we have **medians, sums, minimums, maximums, n** (count of the nonmissing values), **nmiss** (count of the missing values), **stdev, stats** (equivalent to **n**, **means** and **stdev**), and **data** (lists the data for each cell). In addition, there is a subcommand **proportion** with the syntax

proportion = V E_1;

which gives the proportion of cases that have the value V in column E_1.

12.2 One-Way Analysis of Variance

The data in the table below arose from a study of reading comprehension designed to compare three methods of instruction called basal, DRTA, and strategies. The data comprise scores on a test attained by children receiving each of the methods of instruction. There are 22 observations in each group. This study was conducted by Baumann and Jones of the Purdue School of Education.

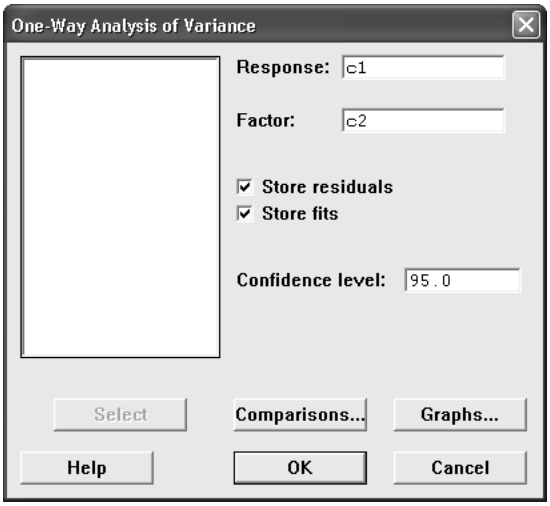
Method	Scores
Basal	4 6 9 12 16 15 14 12 12 8 13 9 12 12 12 10 8 12 11 8 7 9
DRTA	7 7 12 10 16 15 9 8 13 12 7 6 8 9 9 8 9 13 10 8 8 10
Strat	11 7 4 7 7 6 11 14 13 9 12 13 4 13 6 12 6 11 14 8 5 8

We now carry out a one-way analysis of variance on this data to determine if there is any difference between the mean performances of students exposed to the three teaching methods. For this, we use the Stat ▶ ANOVA ▶ One-way command. For this example, there are $I = 3$ levels corresponding to the values Basal, DRTA, and Strat and $n_1 = n_2 = n_3 = 22$. Suppose that we have the values of the x_{ij} in C1 and the corresponding values of the categorical explanatory variable in C2, where Basal is indicated by `1`, DRTA by `2`, and Strat by `3`. The Stat ▶ ANOVA ▶ One-way command together with the dialog boxes shown in Displays 12.2.1, 12.2.2, and 12.2.3 (described below) produce the output

```
One-way ANOVA: C1 versus C2
Source     DF      SS      MS     F      P
C2          2   20.58   10.29  1.13  0.329
Error      63  572.45    9.09
Total      65  593.03

S = 3.014 R-Sq = 3.47% R-Sq(adj) = 0.41%

                     Individual 95% CIs For Mean Based on
                     Pooled StDev
Level  N   Mean StDev  -+---------+---------+---------+--------
Basal 22 10.500 2.972       (-------------*-------------)
DRTA  22  9.727 2.694    (-------------*-------------)
Strat 22  9.136 3.342   (-----------*-----------)
                        -+---------+---------+---------+--------
                        8.0       9.0      10.0      11.0
Pooled StDev = 3.014     8.4 9.6 10.8 12.0

Fisher 95% Individual Confidence Intervals
All Pairwise Comparisons among Levels of C2
Simultaneous confidence level = 87.90%

C2 = Basal subtracted from:
C2      Lower  Center  Upper  -+---------+---------+---------+
DRTA   -2.589  -0.773  1.044    (-----------*-----------)
Strat  -3.180  -1.364  0.453   (-----------*-----------)
                              -+---------+---------+---------+
                             -3.0      -1.5       0.0       1.5

C2 = DRTA subtracted from:
C2      Lower  Center  Upper  -+---------+---------+---------+
Strat  -2.407  -0.591  1.225  (-----------*-----------)
                               -+---------+---------+---------+
                             -3.0      -1.5       0.0       1.5
```

in the Session window. The F test in the ANOVA table with a P-value of 0.329 indicates that the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ would not be rejected. Also, the estimate of σ is given by $s = 3.014$ and 95% confidence intervals are plotted for the individual μ_i.

The dialog box of Display 12.2.1 carries out a one-way ANOVA for the data in C1, with the levels in C2, and puts the ordinary residuals in a variable called `RESI1` and the fitted values in a variable called `FITS1`. Note that because we assume a constant standard deviation and the number of observations is the same in each group, the ordinary residuals can be used in place of standardized residuals. Note also that the ith fitted value in this case is given by the mean of the group to which the observation belongs.



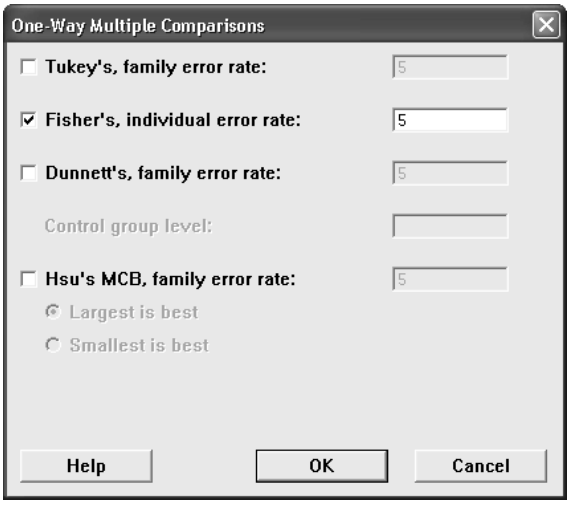
Display 12.2.1: Dialog box for one-way ANOVA.

The dialog box of Display 12.2.2 is obtained by clicking on the Comparisons button in the dialog box of Display 12.2.1. We use this dialog box to select a multiple comparison procedure. Here we have chosen to use the Fisher multiple comparison method with an individual error rate on the comparisons of 5%. This gives confidence intervals for the differences between the means using

$$\bar{y}_i - \bar{y}_j \pm s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\, t^*$$

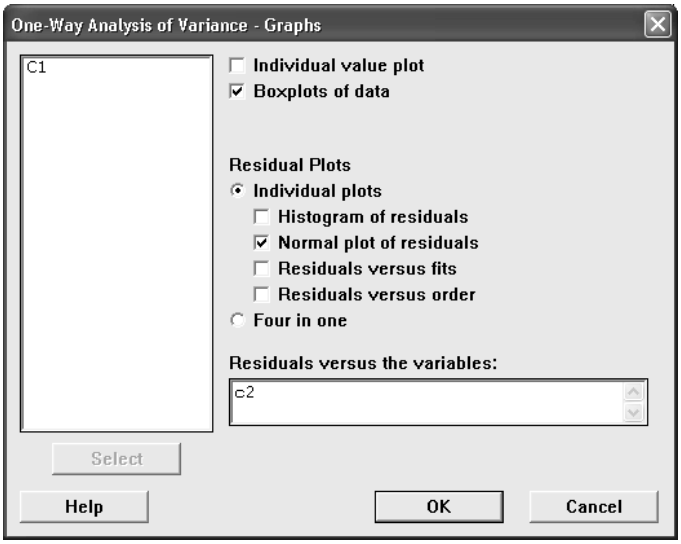
where s is the pooled standard deviation and t^* is the 0.975 percentile of the Student distribution with the error degrees of freedom. Note that with an individual 95% confidence interval, the probability of not covering the true difference (the *individual error rate*) is .05 but the probability of at least one of these three not covering the difference (the *family error rate*) is $1 - .879 = 0.121$. If you want a more conservative family error rate, specify a lower individual error rate. For example, an individual error rate of 0.02 specifies a family error rate of 0.0516 in this example. We refer the reader to Help for details on the other available multiple comparison procedures. In the output above, we see that a 95% confidence interval for $\mu_1 - \mu_2$ is given by $(-1.043, 2.589)$, and because this includes 0, we conclude that there is no evidence against the null hypothesis $H_0 : \mu_1 = \mu_2$. We get the same result for the other two comparisons. Given that the F test

has already concluded that there is no evidence of any differences among the means, there is no reason for us to carry out these individual comparisons, and we do it only for illustration purposes here.

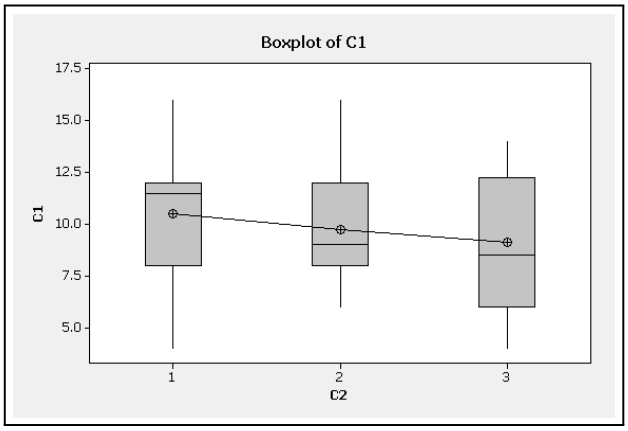


Display 12.2.2: Dialog box for selecting a multiple comparison procedure in a one-way ANOVA.

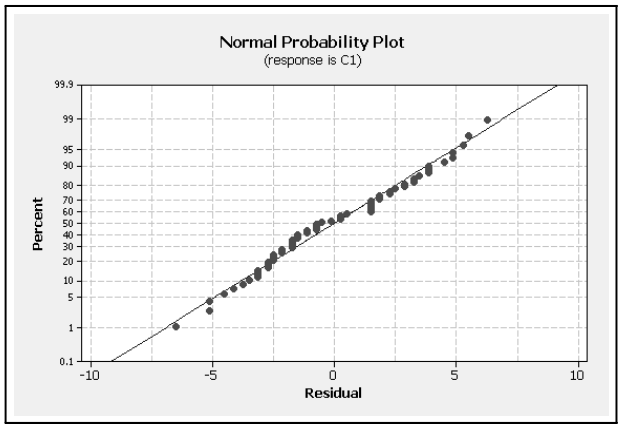
The dialog box of Display 12.2.3 is obtained by clicking on the Graphs button in the dialog box of Display 12.2.1. We have requested a plot of side-by-side boxplots of the data by level, which results in Display 12.2.4, the normal probability plot of the residuals that appears in Display 12.2.5 and a plot of the residuals against the index in C2 that appears in Display 12.2.6. The residual plots don't indicate any problems with the model assumptions.



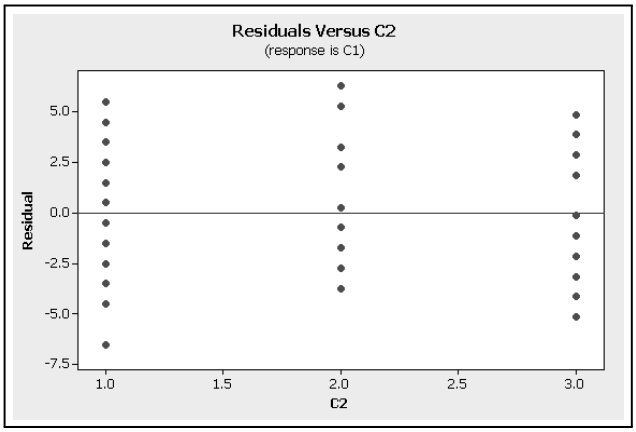
Display 12.2.3: Dialog box for producing plots in a one-way ANOVA.



Display 12.2.4: Boxplots for the example.



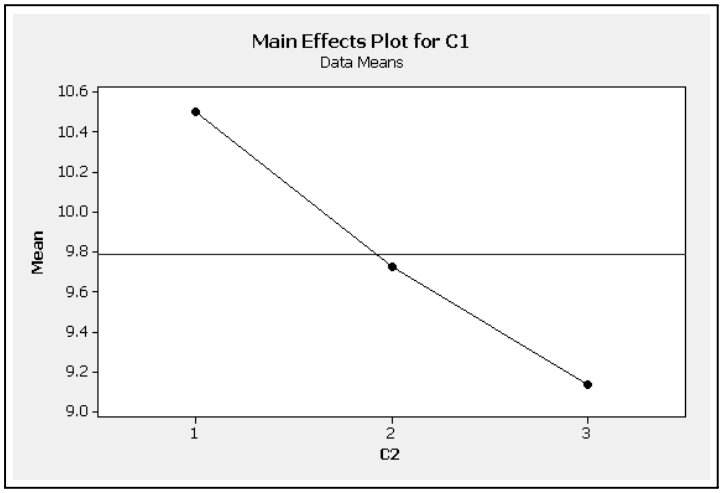
Display 12.2.5: Normal probability plot for the example of this section after fitting a one-way ANOVA model.



Display 12.2.6: Plot of residuals against level for the example of this section after fitting a one-way ANOVA model.

A one-way ANOVA can also be carried out using Stat ▶ ANOVA ▶ One-way (Unstacked) and filling in the dialog box appropriately. This command is much more limited in its features than Stat ▶ ANOVA ▶ One-way, however. So if you have a worksheet with the samples for each level in columns, it would seem better in general to use the Data ▶ Stack command to place the data in one column and then use Stat ▶ ANOVA ▶ One-way.

Also available are analysis of means (ANOM) plots via Stat ▶ ANOVA ▶ Analysis of Means (see Help for details on these) and plots of the means with error bars (± one standard error of the observations at a level) via Stat ▶ ANOVA ▶ Interval Plot. Further, we can plot the means joined by lines using Stat ▶ ANOVA ▶ Main Effects plots as in Display 12.2.7. The dotted line is the grand mean. Power calculations can be carried out using Stat ▶ Power and Sample Size ▶ One-way ANOVA and filling in the dialog box appropriately.



Display 12.2.7: Main effects plot for the example of this section.

The corresponding session command is given by **onewayaov** and has the general syntax

onewayaov E_1 E_2 E_3 E_4

where E_1 is a variable containing the responses, E_2 is a variable containing indices that indicate group membership, E_3 is a variable to hold the residuals, and E_4 is a variable to hold the fitted values. Of course, E_3 and E_4 can be dropped if they are not needed. There are various subcommands that can be used. The **gboxplot** subcommand produces side-by-side boxplots. The **gnormal** subcommand produces a normal probability plot of the residuals. The **gvariables** E_1 subcommand results in a plot of the residuals against the variable E_1. We could also obtain side-by-side dotplots of the data using the **gdotplot** subcommand, a histogram of the residuals using the **ghistogram** subcommand, a plot of the residuals against observation order using the **gorder** subcommand, and a plot of the residuals against the fitted values using the **gfits** subcommand. The **fisher** V_1 subcommand gives confidence intervals for the differences between

the means, where V_1 is the individual error rate. Also available for multiple comparisons are the **tukey, dunnett,** and **mcb** subcommands. For example, the commands

```
MTB > onewayaov c1 c2 c3 c4;
SUBC> gboxplot;
SUBC> gnormal;
SUBC> gvariable c2;
SUBC> fisher.
```

result in the same output as we produced for the example of this section using the menu commands. Here the fits are stored in C4 and the residuals are stored in C3.

The **aovoneway** command can be used for a one-way ANOVA when the data for each level is in a separate column. For example, suppose that the three samples for the example of this section are in columns C3–C5. Then the command

```
MTB > aovoneway c3-c5
```

produces the same ANOVA table and confidence intervals for the means as **onewayaov.** Only a limited number of subcommands are available with this command, however.

12.3 Exercises

1. Generate a sample of 10 from each of the $N(\mu_i, \sigma)$ distributions for $i = 1, \ldots, 5$, where $\mu_1 = 1, \mu_2 = 1, \mu_3 = 1, \mu_4 = 1, \mu_5 = 2$, and $\sigma = 3$. Carry out a one-way ANOVA and produce a normal probability plot of the residuals and the residuals against the explanatory variable. Compute .95 confidence intervals for the differences between the means. Compute an approximate set of .95 simultaneous confidence intervals for the differences between the means.

2. Generate a sample of 10 from each of the $N(\mu_i, \sigma_i)$ distributions for $i = 1, \ldots, 5$, where $\mu_1 = 1, \mu_2 = 1, \mu_3 = 1, \mu_4 = 1, \mu_5 = 2$, $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 3$, and $\sigma_5 = 8$. Carry out a one-way ANOVA and produce a normal probability plot of the residuals and the residuals against the explanatory variable. Compare the residual plots with those obtained in Exercise II.12.1.

3. The F statistic in a one-way ANOVA, when the standard deviation σ is constant from one level to another, is distributed *noncentral* $F(k_1, k_2)$ with noncentrality λ, where $k_1 = I - 1$, $k_2 = n_1 + \cdots n_I - I$,

$$\lambda = \frac{\sum_{i=1}^{I} n_i \left(\mu_i - \bar{\mu}\right)^2}{\sigma^2}$$

and $\bar{\mu} = \sum_{i=1}^{I} n_i \mu_i / \sum_{i=1}^{I} n_i$. Using simulation, approximate the power of the test in Exercise II.12.1 with level .05 and the values of the parameters

specified and compare your results with exact results obtained from Stat ▶ Power and Sample Size ▶ One-way ANOVA.

Chapter 13

Two-Way Analysis of Variance

New Minitab command discussed in this chapter

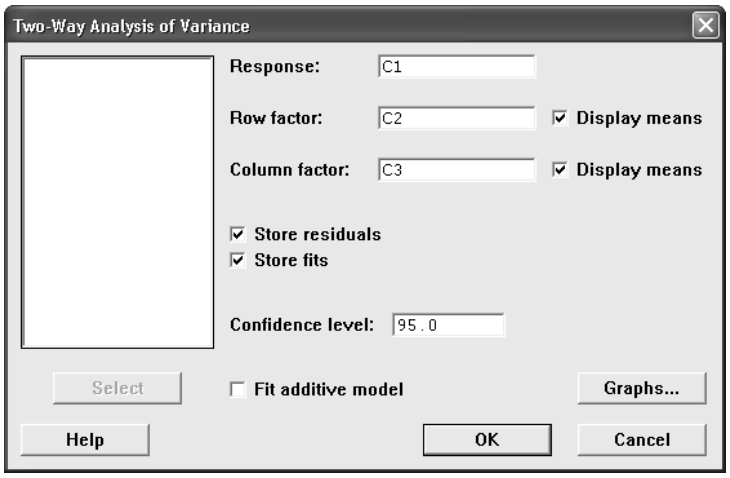
Stat ▶ ANOVA ▶ Two-way

This chapter deals with methods for making inferences about the relationship existing between a single numeric response variable and two categorical explanatory variables. The Stat ▶ ANOVA ▶ Two-way command is used to carry out a two-way ANOVA.

We write the two-way ANOVA model as $x_{ijk} = \mu_{ij} + \epsilon_{ijk}$, where $i = 1, \ldots, I$ and $j = 1, \ldots, J$ index the levels of the categorical explanatory variables and $k = 1, \ldots, n_{ij}$ indexes the individual observations at each treatment (combination of levels), μ_{ij} is the mean response at the *ith* level and the *jth* level of the first and second explanatory variable, respectively, and the errors ϵ_{ijk} are a sample from the $N(0, \sigma)$ distribution. Based on the observed x_{ijk}, we want to make inferences about the unknown values of the parameters $\mu_{11}, \ldots, \mu_{IJ}, \sigma$.

13.1 The Two-Way ANOVA Command

We consider a generated example, where $I = J = 2, \mu_{11} = \mu_{21} = \mu_{12} = \mu_{22} = 1, \sigma = 2$, and $n_{11} = n_{21} = n_{12} = n_{22} = 5$. The ϵ_{ijk} are generated as a sample from the $N(0, 2)$ distribution, and we put $x_{ijk} = \mu_{ij} + \epsilon_{ijk}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ and $k = 1, \ldots, n_{ij}$. Note that the Stat ▶ ANOVA ▶ Two-way command requires balanced data; i.e., all the n_{ij} must be equal. We pretend that we don't know the values of the parameters and carry out a two-way analysis of variance. If the x_{ijk} are in C1, the values of i in C2 and the values of j in C3, the dialog box of Display 13.1.1



Display 13.1.1: Dialog box for producing a two-way analysis of variance.

produces the following output.

```
Two-way ANOVA: C1 versus C2, C3
Source      DF       SS       MS     F      P
C2           1  11.8566  11.8566  3.28  0.089
C3           1   0.0737   0.0737  0.02  0.888
Interaction  1   9.9479   9.9479  2.75  0.117
Error       16  57.9188   3.6199
Total       19  79.7970

S = 1.903 R-Sq = 27.42% R-Sq(adj) = 13.81%

              Individual 95% CIs For Mean Based on
              Pooled StDev
C2    Mean    -+---------+---------+---------+--------
1  2.72136          (----------*---------)
2  1.18146    (----------*---------)
              -+---------+---------+---------+--------
              0.0       1.2       2.4       3.6

              Individual 95% CIs For Mean Based on
              Pooled StDev
C3    Mean    -+---------+---------+---------+--------
1  1.89070    (-----------------*-----------------)
2  2.01212     (-----------------*-----------------)
              -+---------+---------+---------+--------
              0.70      1.40      2.10      2.80
```

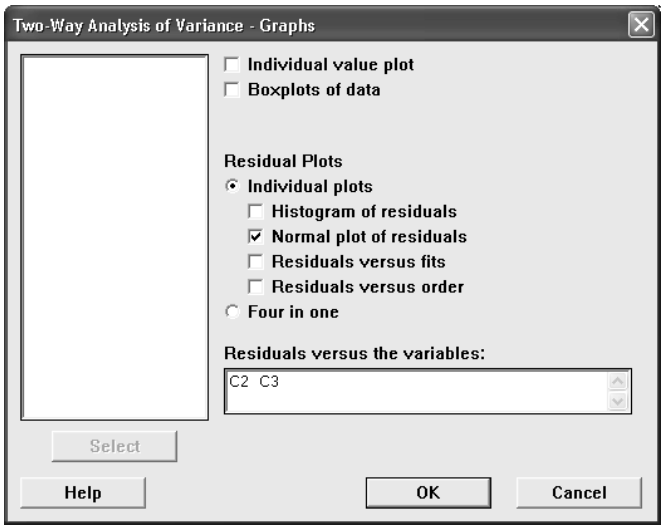
We see from this that the null hypothesis of no interaction is not rejected (P-value = .117) and neither is the null hypothesis of no effect due to the `C2` factor (P-value = .089) nor the null hypothesis of no effect due to factor C3 (P-value = .888) as is appropriate.

Note that by checking the Display means boxes in the dialog box of Display 13.1.1 we have caused 95% confidence intervals to be printed for the response means at each value of C2 and each value of C3, respectively. These cell means are relevant only when we decide that there is no interaction, as is the case here, and we note that all the intervals contain the true value 1 of these means.

We also checked the Store residuals and Store fits in the dialog boxes of Display 13.1.1. This results in the (ordinary) residuals being stored in C4 and the fitted values (cell means) being stored in C5. If these columns already had entries the next two available columns would be used instead.

If we want to fit the model without any interaction, supposing we know this to be true, we can check the Fit additive model box in the dialog box of Display 13.1.1. This is acceptable only in rare circumstances, however, as it is unlikely that we will know that this is true.

Various graphs are also available via the Graphs button in the dialog box of Display 13.1.1. Clicking on this results in the Dialog box shown in Display 13.1.2. Here we have asked for a normal probability plot of the (ordinary) residuals and a plot of the (ordinary) residuals versus the variables C2 and C3. Recall that with balance it is acceptable to use the ordinary residuals rather than the standardized residuals. We haven't reproduced the corresponding plots here but, as we might expect, they gave no grounds for suspecting the correctness of the model.

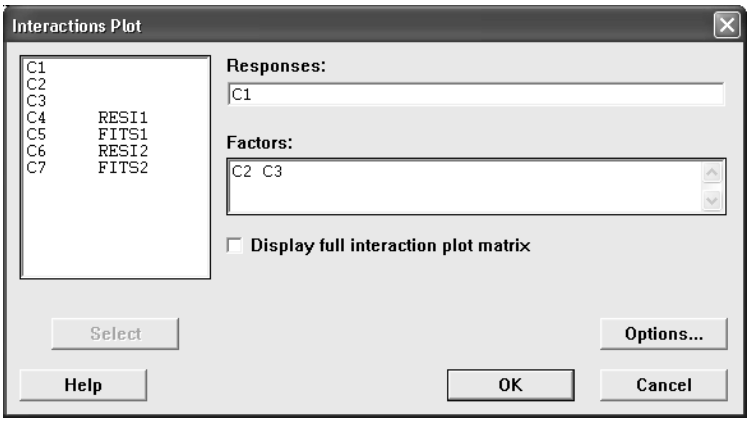


Display 13.1.2: Dialog box for producing various residual plots obtained via the Graphs button in the dialog box of Display 13.1.1.

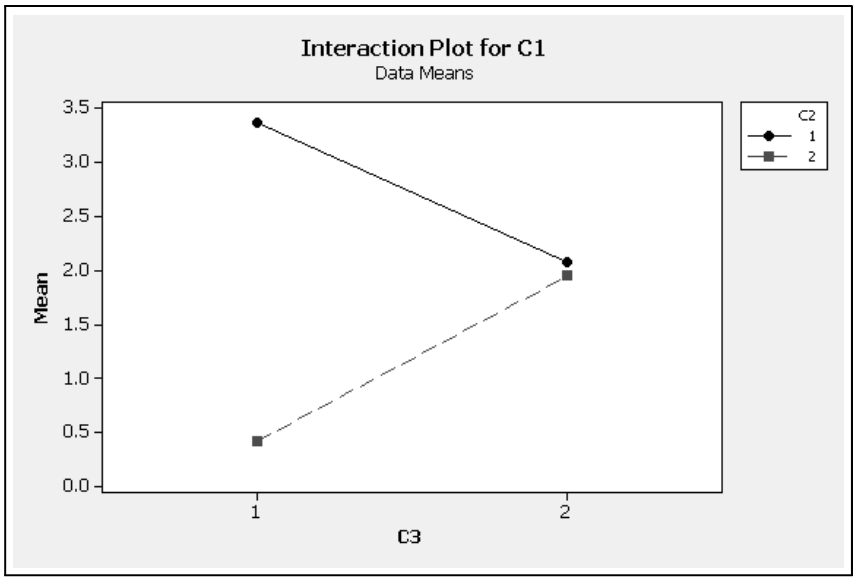
If we conclude that there is an interaction then we must look at the individual IJ cell means to determine where the interaction occurs. A plot of these cell means is often useful in this regard. Also available are analysis of means (ANOM) plots via Stat ▶ ANOVA ▶ Analysis of Means. In addition, we can plot the marginal means joined by lines using Stat ▶ ANOVA ▶ Main Effects

Plot and plot the cell means joined by lines using Stat ▶ ANOVA ▶ Interaction Plot using the dialog box of Display 13.1.3 with the output in Display 13.1.4.

Note that while the plot seems to indicate an interaction, this is not confirmed by the statistical test. Power calculations can be carried out using Stat ▶ Power and Sample Size ▶ 2-Level Factorial Design and filling in the dialog box appropriately. Commands are available in Minitab for analyzing unbalanced data and for situations where there are more than two factors where some factors are continuous and some categorical, and so on.



Display 13.1.3: Dialog box for obtaining the interaction plot of Display 13.1.4.



Display 13.1.4: Plot of cell means in two-way ANOVA simulated example.

The corresponding session command for carrying out a two-way ANOVA is given by **twowayaov.** For example, the command

```
MTB > twowayaov c1 c2 c3 c4 c5;
SUBC> gnormal;
SUBC> gvariable c2 c3;
SUBC> means c2 c3.'
```

results in the same output as above. The **gnormal** subcommand results in a normal probability plot of the residuals being plotted while the **gvariables** subcommand results in a plot of the residuals against each of the factors C2 and C3. The **ghistogram, gfits,** and **gorder** subcommands are also available for a histogram of the residuals, the residuals against the fitted values, and the residuals against observation order, respectively. The **means** subcommand causes the estimates of marginal means for each level of C2 and C3 to be printed together with 95% confidence intervals. If we want to fit the model without any interaction, supposing we know this to be true, then the **additive** subcommand is available to do this.

13.2 Exercises

1. Suppose $I = J = 2, \mu_{11} = \mu_{21} = 1$, $\mu_{12} = \mu_{22} = 2, \sigma = 2$, and $n_{11} = n_{21} = n_{12} = n_{22} = 10$. Generate the data for this situation, and carry out a two-way analysis. Plot the cell means (an interaction effect plot). Do your conclusions agree with what you know to be true?

2. Suppose $I = J = 2, \mu_{11} = \mu_{21} = 1$, $\mu_{12} = 3, \mu_{22} = 2, \sigma = 2$, and $n_{11} = n_{21} = n_{12} = n_{22} = 10$. Generate the data for this situation, and carry out a two-way analysis. Plot the cell means (an interaction effect plot). Do your conclusions agree with what you know to be true?

3. Suppose $I = J = 2, \mu_{11} = \mu_{21} = 1$, $\mu_{12} = \mu_{22} = 2, \sigma = 2$, and $n_{11} = n_{21} = n_{12} = n_{22} = 10$. Generate the data for this situation, and carry out a two-way analysis. Form 95% confidence intervals for the marginal means. Repeat your analysis using the additive model and compare the confidence intervals. Can you explain your results?

Chapter 14

Bootstrap Methods and Permutation Tests

This chapter is concerned with computationally intensive inference methods that are sometimes applicable when methods based on strong assumptions, such as normality, cannot be used because it is clear that the assumptions are not satisfied. These methods are based on repeated sampling from a column of fixed data. Bootstrap sampling requires that we sample this column with replacement, and permutation tests require that we sample the column without replacement. In the next sections we describe how to use Minitab to accomplish this.

At this point Minitab does not have built-in commands to implement bootstrap sampling or permutation tests. For this we need some of the programming features of Minitab. Actually you will not have to learn how to program as we will provide the necessary code and explain how to use it in the following sections. It is a simple matter to modify this code so that different statistics can be used.

A Minitab program is called a *macro* and must start with the statement `gmacro` and end with the statement `endmacro`. The first statement after `gmacro` gives a name to the program. Comments in a program, put there for explanatory purposes, start with `note.`

If the file containing the program is called `prog.txt` and this is stored in the root directory of a disk drive called `c,` then the Minitab command

```
MTB> %c:/prog.txt
```

will run the program. Any output will either be printed in the Session window (if you have used a `print` command) or stored in the Minitab worksheet. Basically, this is all you need to know to run the programs discussed in this chapter.

14.1 Bootstrap Sampling

Suppose the data in the following table of $n = 15$ values is stored in C1 and we wish to calculate the bootstrap distribution of the sample median that we are using to estimate the mean of the population distribution.

0.2	3.0	2.2	1.0	4.0
0.5	2.3	−1.3	3.1	−1.0
5.8	0.4	1.3	−2.7	−8.6

The sample median for this data is given by 1.00.

The following Minitab code generates 1000 bootstrap samples from the data in C1, calculates the median of each of these samples, and then calculates the sample mean and variance of these medians.

```
gmacro
bootstrapping
base 34256734
note - original sample is stored in c1
note - bootstrap sample is placed in c2 (each one overwritten)
note - medians of bootstrap samples are stored in c3
note - k1 = size of data set (and bootstrap samples)
let k1=15
do k2=1:1000
note - the upper bound for k2 = the number of bootstrap
note - samples generated, here this is 1000 and can be changed
sample 15 c1 c2;
replace.
note - you must replace the following line with the Minitab
note   commands for whatever statistic you want to bootstrap
let c3(k2)=median(c2)
enddo
note - k3 equals the mean of the bootstrapped median
let k3=mean(c3)
note - k4 equals the sample variance of the bootstrapped median
let k4=(stdev(c3))**2
print 'bootstrap mean' k3 'bootstrap variance' k4
endmacro
```

To change the number of bootstrap samples we generate we must change the ninth line. Currently it reads

```
do k2=1:1000
```

so that we are generating 1000 bootstrap samples. If we want to generate 10,000 bootstrap samples, then we must change this to

```
do k2=1:10000
```

and of course any other number can be substituted. Be careful though, as the bigger we choose this number the longer we have to wait for the computations to be carried out.

The entire bootstrap sample of medians is stored in C3. So we can plot this in a histogram to get some idea of what the distribution looks like that the bootstrap procedure is sampling from.

We put the above code in a file `bootstrap.txt` and stored this in the main directory of the `c` drive. Then the command

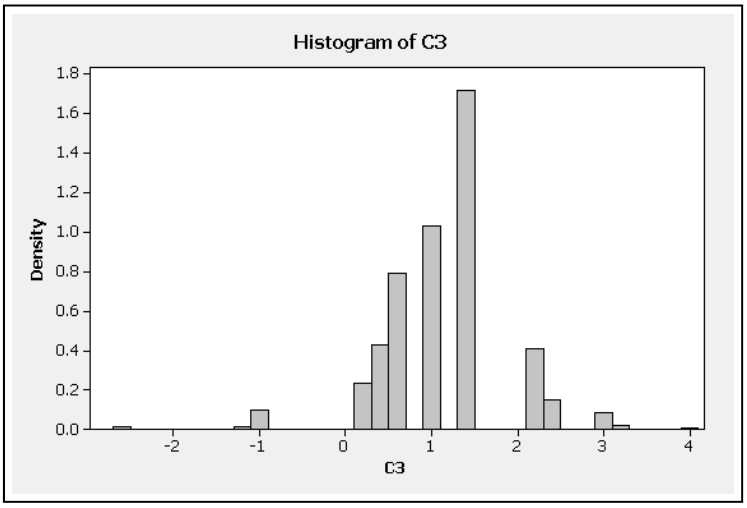
```
MTB > %c:/bootstrap.txt
```

runs these commands and produces the output

```
bootstrap mean
K3 1.06350
bootstrap variance
K4 0.514652
```

which gives the estimate bootstrap mean and bootstrap variance as 1.06350 and 0.514652, respectively. So the bias is $1.06350 - 1.00 = 0.0635$, which is relatively small.

Using the Graph ▶ Histogram command on the values stored in C3 we produced the plot in Display 14.1.1. We can see from this that the bootstrap distribution of the median is not very normal looking.



Display 14.1.1: Histogram of 1000 bootstrap sample medians.

There are a number of built-in Minitab functions, such as **median**, whose bootstrap distribution we are often interested in. There are others, however, for which we must do a bit of programming. For example, we must program the various trimmed means. If we want an α-trimmed mean, where $\alpha \in [0, 1]$, then we remove the m smallest observations and the m largest observations from the sample and calculate the mean of the rest, where m is the closest integer to αn.

We now provide an example of obtaining the bootstrap distribution of a 25%-trimmed mean of the data given above. Note that in this case, since $(.25)(15) = 3.75$, we take $m = 4$, and this implies that we remove the observations $-8.6, -2.7, -1.3, 3.1, 4.0$ and 5.8 from the sample. The .25-trimmed mean is then given by 1.10. We then used the following code to estimate the bootstrap distribution of the .25-trimmed mean.

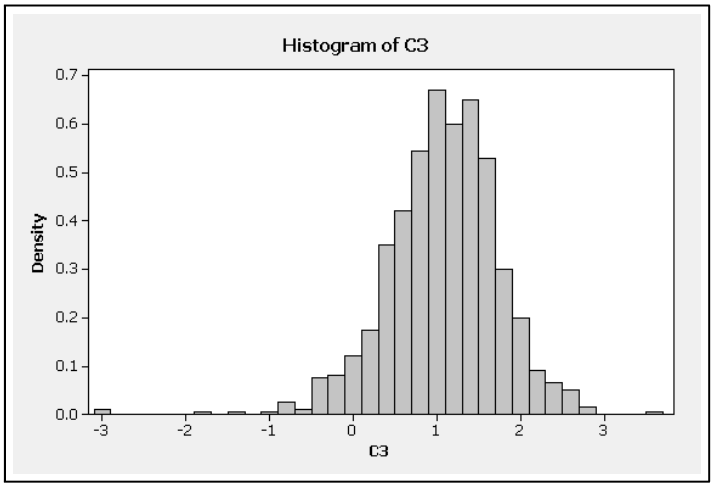
```
gmacro
bootstrapping
base 34256734
note - original sample is stored in c1
note - bootstrap sample is placed in c2 (each one overwritten)
note - the sorted bootstrap sample is then put in c2
note - 25% trimmed means of bootstrap samples are computed and
note - stored in c3 for more analysis
do k2=1:1000
sample 15 c1 c2;
replace.
sort c2 c2
let k4=0
do k3=4:12
let k4=k4+c2(k3)
enddo
let c3(k2)=k4/9
enddo
let k5=mean(c3)
let k6=(stdev(c3))**2
print 'bootstrap mean' k5 'bootstrap variance' k6
endmacro
```

Note that the code in lines 13-18, namely,

```
let k4=0
do k3=4:12
let k4=k4+c2(k3)
enddo
let c3(k2)=k4/9
enddo
```

calculates the .25-trimmed mean for this data and needs to be changed appropriately for other trimmed means and other data sets. Running this program we obtained the estimated mean of the bootstrap distribution as 1.06828 and the estimated bootstrap variance as 0.440809. So in this case the bias is $1.06828 - 1.10 = -0.03172$, which is reasonably small.

Using the Graph ▶ Histogram command on the values stored in C3 we produced the plot in Display 14.1.2. We can see from this that the bootstrap distribution of the median is much more normal looking.



Display 14.1.2: Histogram of 1000 bootstrap sample .25-trimmed means.

Ignoring the skewness of the bootstrap distribution, the bootstrap t .95-confidence interval for the population .25 trimmed mean, is then given by

$$\begin{aligned} 1.10 \pm t_{.975}(14)\sqrt{0.531869} &= 1.10 \mp (2.14479)\sqrt{0.440809} \\ &= [-0.324, 2.524]. \end{aligned}$$

To calculate the *bootstrap percentile confidence intervals* we first sort the bootstrap distribution values in C3 and find the .025 and the .975 percentiles of this sample. The commands

```
MTB > sort c3 c4
MTB > set c5
DATA> 1:1000
DATA> end
MTB > let c5=c5/1000
```

place the sorted values in C4 and then calculates the proportion of values less than or equal to each value and places these proportions in C5. We then record the values in C4 that correspond to .025 and .975 in C5. In this case we obtained $(-0.34444, 2.31111)$ as the .95-bootstrap percentile confidence interval. We note that this interval is similar to the bootstrap t interval.

14.2 Permutation Tests

As with bootstrapping Minitab does not have built-in commands to carry out permutation tests. Again, however, it is very easy to program Minitab to implement these tests.

We illustrate how to implement a permutation test using a data set where we have two samples, one from a treatment and the other from a control. In

the following table, T stands for Treatment and C for Control.

T	T	T	T	C	C	C	C
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	

We want to test the null hypothesis that the mean of the distribution for the treatment group is the same as the mean of the distribution for the control group. Now suppose that we have the values stored in C2 with an index stored in C1 that indicates whether the value is from the Treatment group or from the Control group. Then the commands

```
MTB > unstack c2 c3 c4;
SUBC> subscripts c1.
MTB > let k1=mean(c3)
MTB > let k2=mean(c4)
MTB > let k3=k1-k2
MTB > print k3
Data Display
K3 9.95445
```

calculate the means of the T group and the C group, the difference of the two means and then prints this quantity. We obtain 9.95445 as the difference of the means.

The following commands compute the P-value based on permutation distribution of the difference of means to test the null hypothesis that the means of the T and C groups are the same against the alternative that the mean of the T group is greater than the mean of the C group

```
gmacro
permutation
base 468798
note - index is stored in c1
note - original samples are stored in c2
note - the following commands compute the difference of the
note - means for the original samples
note - and stores this difference in k10
unstack c2 c4 c5;
subscripts c1.
let k2=mean(c4)
let k3=mean(c5)
let k10=k2-k3
note - permuted samples are stored in c3
note - unstacked permuted samples are stored in c4 and c5
```

```
note - the difference in means is stored in c6
note - the value 1 is stored in c7 if difference in means of
note - these samples is greater than k10 and the value 0 is
note - stored there otherwise
do k1=1:1000
sample 44 c2 c3
unstack c3 c4 c5;
subscripts c1.
let k2=mean(c4)
let k3=mean(c5)
let k4=k2-k3
let c6(k1) = k4
let c7(k1) = k4 >= k10
enddo
note - the mean of c7 is the proportion of the differences of
note - means in the permutation distribution that are greater
note - than or equal to the observed difference
let k5=mean(c7)
print k10 k5
endmacro.
```

The output from the above program is

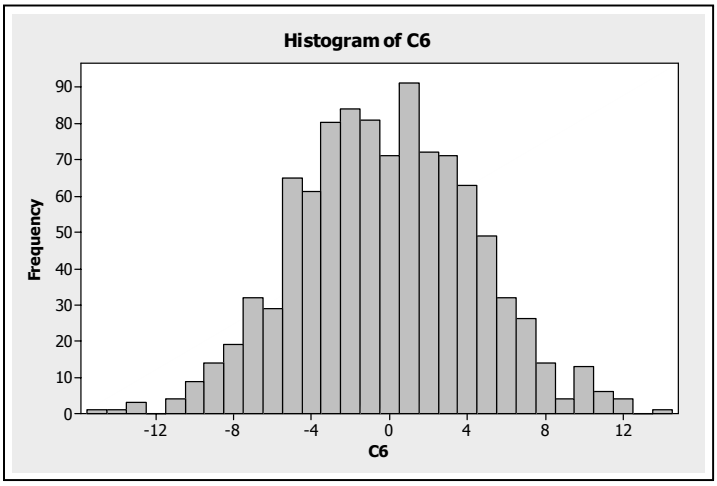
```
K10 9.95445
K5 0.0210000
```

and this tells us that the P-value is .021, and so we can conclude that we have evidence against the null hypothesis.

Note that the above program stores the sample from the permutation distribution in C6 so we can analyze this further. For example, Display 14.2.1 gives a histogram of the 1000 differences of means as obtained in the above program. We see that this is reasonably normal looking.

A two-sided permutation test can be carried out in this case by simply computing the proportion of differences that are greater in absolute value than the absolute value of the observed difference, which in this case equals $|9.95445| = 9.95445$. The following commands accomplish this.

```
do k1=1:1000
sample 44 c2 c3
unstack c3 c4 c5;
subscripts c1.
let k2=mean(c4)
let k3=mean(c5)
let k4=k2-k3
let c6(k1) = k4
let c7(k1) = abs(k4) >= abs(k10)
enddo
let k5=mean(c7)
```



Display 14.2.1: Histogram of 1000 differences of means obtained by randomly permuting the samples.

This produced the output

```
K5 0.0330000
```

so the results are significant when using the two-sided alternative as well.

For the matched pairs permutation test for comparing treatment A to treatment B we randomly assign an individual's A measurement to A or B, and the B measurement is assigned the other label. We then compare the observed mean difference with the distribution of these differences obtained from all possible random assignments. The following code carries out the two-sided matched pair permutation test when we have 10 observations with the A measurements stored in C1 and the B measurements stored in C2.

```
gmacro
permutationmatched
base 468798
note - first measurement is stored in c1
note - second measurement is stored in c2
note - differences stored in c3
note - k2 = observed mean difference
let c3=c1-c2
let k2=mean(c3)
note - randomly choose which observations in c1 will be
note - labelled A (10 values generated from Bernoulli(.5))
note - whenever a 1 occurs in c4 multiply entry in c3 by 1
note - otherwise multiply by -1, store in c6
note - and put mean difference in k4 and store in c7
do k1=1:1000
random 10 c4;
bernoulli .5.
```

```
let c5=-1+2*c4
let c6=c5*c3
let k3=mean(c6)
let c7(k1) = k3
let c8(k1) = abs(k3) >= abs(k2)
enddo
let k4=mean(c8)
print k2 k4
endmacro
```

14.3 Exercises

1. Generate a sample of $n = 20$ from the $N(0, 1)$ distribution. Approximate the bootstrap distribution of $\bar{x}$ by generating 1000 bootstrap samples. Estimate the bias, estimate the bootstrap variance, and plot the 1000 values of the sample mean in a density histogram. Calculate, and compare, .95 confidence intervals for the population mean based on the t distribution and bootstrap distribution.

2. Generate a sample of $n = 20$ from the Chi-squared(1) distribution. Approximate the bootstrap distribution of $\bar{x}$ by generating 1000 bootstrap samples. Estimate the bias, estimate the bootstrap variance, and plot the 1000 values of the sample mean in a density histogram. Calculate, and compare, .95 confidence intervals for the population mean based on the t distribution and bootstrap distribution.

3. Generate a sample of $n = 20$ from the $N(0, 1)$ distribution. Approximate the bootstrap distribution of the .1-trimmed mean by generating 1000 bootstrap samples. Estimate the bias, estimate the bootstrap variance, and plot the 1000 values of the .1-trimmed mean in a density histogram. Calculate, and compare, .95 confidence intervals for the population .1-trimmed mean based on the t distribution and bootstrap distribution.

4. Generate a sample of $n = 20$ from the Chi-squared(1) distribution. Approximate the bootstrap distribution of the .1-trimmed mean by generating 1000 bootstrap samples. Estimate the bias, estimate the bootstrap variance, and plot the 1000 values of the .1-trimmed mean in a density histogram. Calculate, and compare, .95 confidence intervals for the population .1-trimmed mean based on the t distribution and bootstrap distribution.

5. Generate a sample of 10 from the $N(0, 1)$ distribution and a sample of 15 from the $N(2, 1)$ distribution and carry out a two-sided permutation test that the difference of means is 0. Compare the P-value obtained with that obtained from a two-sided t test.

6. Generate a sample of 10 from the Student(1) distribution and a sample of 15 from the Student(1) + 2 distribution (generate a sample from the Student(1) and add 2 to each sample element) and carry out a two-sided permutation test that the difference of means is 0. Compare the P-value obtained with that obtained from a two-sided t test.

7. Generate a sample of 10 from the $N(0, 1)$ distribution and a sample of 10 from the $N(2, 1)$ distribution and carry out a two-sided matched pair permutation test that the difference of means is 0. Compare the P-value obtained with that obtained from a two-sided matched pair t test.

Chapter 15

Nonparametric Tests

New Minitab commands discussed in this chapter

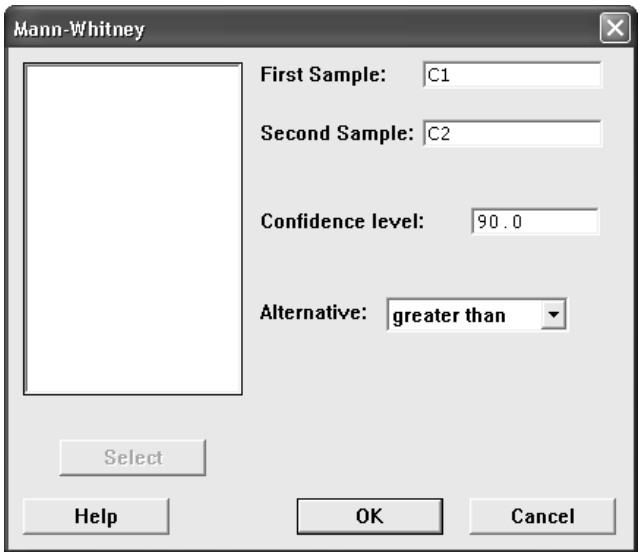
Stat ▶ Nonparametrics ▶ Kruskal-Wallis
Stat ▶ Nonparametrics ▶ Mann-Whitney
Stat ▶ Nonparametrics ▶ 1-Sample Wilcoxon

This chapter deals with inference methods that do not depend upon the assumption of normality. These methods are sometimes called *nonparametric* or *distribution free* methods. Recall that we discussed a distribution-free method in Section 7.4, where we presented the Stat ▶ Nonparametrics ▶ 1-Sample Sign command for the sign confidence interval and sign test for the median. Recall also the Data ▶ Rank command in I.10.6, which can be used to compute the ranks of a data set.

15.1 The Wilcoxon Rank Sum Procedures

The Mann-Whitney test for a difference between the locations of two distributions is equivalent to the Wilcoxon rank sum test in the following sense. Suppose that we have two independent samples $y_{11}, \ldots, y_{1n_1}$ and $y_{21}, \ldots, y_{2n_2}$ from two distributions that differ at most in their locations as represented by their medians. The Mann-Whitney statistic U is the number of pairs (y_{1i}, y_{2j}) where $y_{1i} > y_{2j}$, while the Wilcoxon rank sum test statistic W is the sum of the ranks from the first sample when the ranks are computed for the two samples considered as one sample combined. It can be shown that $W = U + n_1(n_{1+1})/2$ and so the test procedures based on these statistics are equivalent.

Suppose we have one sample of four values 166.7, 172.2, 165.0, and 176.9 stored in C1 and a second sample of four values 158.6, 176.4, 153.1, and 156.0 stored in C2. The Stat ▶ Nonparametrics ▶ Mann-Whitney command, implemented as in the dialog box of Display 15.1.1,



Display 15.1.1: Dialog box for implementing the Mann-Whitney command.

leads to the output

```
Mann-Whitney Test and CI: C1, C2
    N    Median
C1  4    169.45
C2  4    157.30
Point estimate for ETA1-ETA2 is 11.30
93.9 Percent CI for ETA1-ETA2 is (-9.70,20.90)
W = 23.0
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0.0970
```

which indicates that the test of H_0 : the medians of the two distributions are identical versus H_a : the median of the first distribution is greater than the median of the second gives a P-value of .0970. Also, an estimate of 11.3 is produced for the difference in the medians, and we asked for a 90% confidence interval for this difference by placing `90` in the Confidence level box. Note that exact confidences cannot be attained due to the discrete distribution followed by the statistic U. The Mann-Whitney test requires the assumption that the two distributions we are sampling from have the same form.

The corresponding session command is given by **mann-whitney.** For example, the command

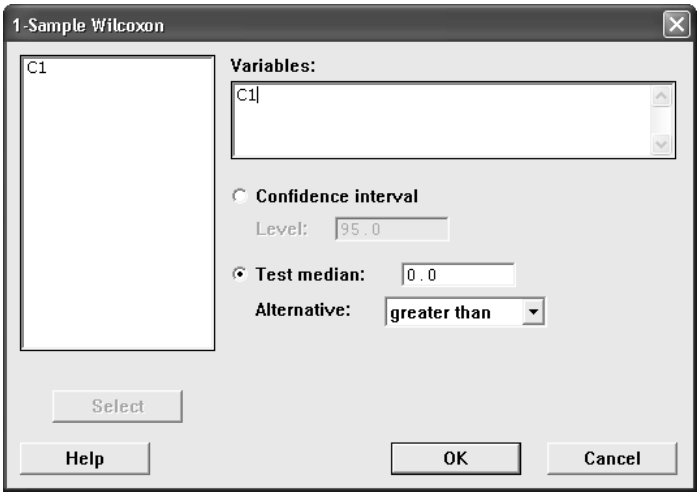
```
MTB > mann-whitney 90 c1 c2;
SUBC> alternative 1.
```

leads to the above output. Note that we have placed 90 on the command line to indicate that we want a 90% confidence interval. If this value is left out, a default 95% confidence interval is computed. Also available are the one-sided test of H_0 : the medians of the two distributions are identical versus H_a :

the median of the first distribution is smaller than the median of the second, using the subcommand **alternative** −1, and the two-sided test is obtained if no **alternative** subcommand is employed.

15.2 The Wilcoxon Signed Rank Procedures

The Wilcoxon signed rank test and confidence interval are used for inferences about the median of a distribution. The Wilcoxon procedures are based on ranks, which is not the case for the sign procedures discussed in Section 7.4. Suppose we have two measurements on each of five individuals. The differences in these measurements are $.37, -.23, .66, -.08, -.17$ and they are stored in C1. The _Stat ▶ Nonparametrics ▶ 1-Sample Wilcoxon command, implemented as in the dialog box in Display 15.2.1,



Display 15.2.1: Dialog box for implementing the Wilcoxon signed rank test.

leads to the output

```
Test of median = 0.000000 versus median > 0.000000
         N for    Wilcoxon         Estimated
      N   Test   Statistic      P     Median
C1    5     5         9.0    0.394    0.1000
```

which gives the P-value .394 for testing H_0 : the median of the difference is 0 versus H_a : the median of the difference is greater than 0. If instead we had filled in the Confidence interval button and placed 90 in the Level box of the dialog box in Display 15.2.1, we would have obtained the output

```
           Estimated     Achieved
      N       Median   Confidence    Confidence Interval
C1    5        0.100         89.4     (-0.200, 0.515)
```

which provides a 90% confidence interval for the median. Note that the Wilcoxon signed rank procedures for the median require an assumption that the response values (in this case the difference) come from a distribution symmetric about its median.

The corresponding session commands are given by **wtest** and **winterval** for tests and confidence intervals respectively. The general syntax of the **wtest** command is

wtest V E_1

where V is the hypothesized value of the median, with 0 being the default value, and E_1 is the column containing the data. For example, the command

```
MTB > wtest c1;
SUBC> alternative 1.
```

produces the above output for the test. The general syntax of the **winterval** command is

winterval V E_1

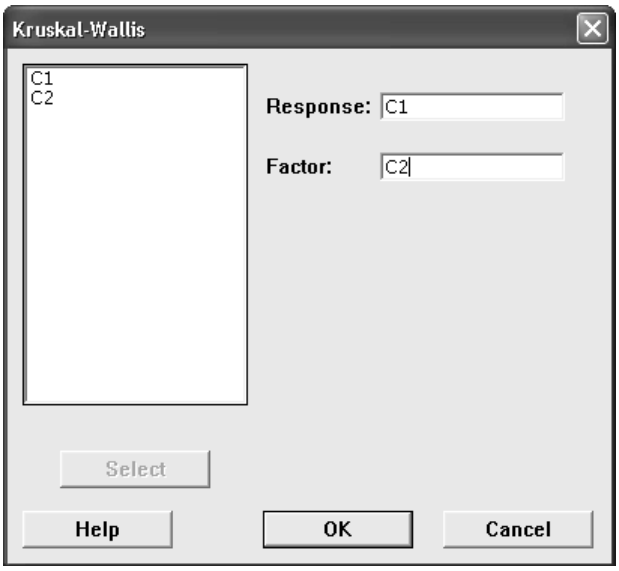
where V is the confidence level, with 0.95 being the default value, and E_1 is the column containing the data.

15.3 The Kruskal-Wallis Test

The Kruskal-Wallis test is the analog of the one-way ANOVA in the nonparametric setting. To illustrate, suppose we use the data in the table of section 12.2 where our purpose is to compare three methods of instruction called basal, DRTA, and strategies. The data comprise scores on a test attained by children receiving each of the methods of instruction. There are 22 observations in each group. We carry out a Kruskal-Wallis test, using this data, to determine if there is any difference between the median performances of students exposed to the three teaching methods. For this there are $I = 3$ levels corresponding to the values Basal, DRTA, and Strat and $n_1 = n_2 = n_3 = 22$. Suppose that we have the scores in C1 and the corresponding values of the categorical explanatory variable in C2, where Basal is indicated by `1`, DRTA by `2`, and Strat by `3`. The Stat ▶ Nonparametrics ▶ Kruskal-Wallis command, as implemented in Display 15.3.1, produces the output

```
Kruskal-Wallis Test:  C1 versus C2
Kruskal-Wallis Test on C1
C2       N   Median  Ave Rank       Z
1       22   11.500      38.1    1.37
2       22    9.000      32.9   -0.19
3       22    8.500      29.6   -1.18
Overall 66               33.5
H = 2.19 DF = 2 P = 0.334
H = 2.22 DF = 2 P = 0.329 (adjusted for ties)
```

which gives a P-value of .334 for testing H_0 : each sample comes from the same distribution versus H_a : at least two of the samples come from different distributions. Note that the validity of the Kruskal-Wallis test relies on the assumption that the distributions being sampled from all have the same form.



Display 15.3.1: Dialog box for implementing the Kruskal-Wallis test.

The corresponding session command is given by **kruskal-wallis.** For example, the command

```
MTB > kruskal-wallis c1 c2
```

also produces the above output. The general syntax of the **kruskal-wallis** command is

kruskal-wallis E_1 E_2

where E_1 contains the data and E_2 contains the levels of the factor.

15.4 Exercises

1. Generate a sample of $n = 10$ from the $N(0, 1)$ distribution and compute the P-value for testing H_0 : the median is 0 versus H_a : the median is not 0, using the t test and the Wilcoxon signed rank test. Compare the P-values. Repeat this with $n = 100$.

2. Generate a sample of $n = 10$ from the $N(0, 1)$ distribution and compute 95% confidence intervals for the median, using the t confidence interval and the Wilcoxon signed rank confidence intervals. Compare the lengths of the confidence intervals. Repeat this with $n = 100$.

3. Generate two samples of $n = 10$ from the Student(1) distribution and add 1 to the second sample. Test H_0 : the medians of the two distributions are identical versus H_a : the medians are not equal using the two-sample t test and using the Mann-Whitney test. Compare the results.

4. Generate a sample of 10 from each of the $N(1, 2), N(2, 2)$, and $N(3, 1)$ distributions. Test for a difference among the distributions using a one-way ANOVA and using the Kruskal-Wallis test. Compare the results.

5. Generate 10 scores for 10 brands from the $N(\mu_{ij}, \sigma)$ distributions for $i = 1, 2$ and $j = 1, 2$, where $\mu_{11} = \mu_{21} = 1$ and $\mu_{12} = \mu_{22} = 2$, and treat each test for no effect due to brand using a two-way ANOVA with the assumption of no interaction and also using the Friedman test. Compare the results.

Chapter 16

Logistic Regression

New Minitab commands discussed in this chapter

Stat ▶ Regression ▶ Binary Logistic Regression
Stat ▶ Regression ▶ Nominal Logistic Regression
Stat ▶ Regression ▶ Ordinal Logistic Regression

This chapter deals with the *logistic regression model.* This model arises when the response variable y is binary—i.e., takes only two values—and we have a number of explanatory variables $x_1, \ldots, x_k$.

16.1 The Logistic Regression Model

The regression techniques discussed in Chapters 10 and 11 require that the response variable y be a continuous variable. In many contexts, however, the response is discrete and in fact binary, i.e., taking the values 0 and 1. Let p denote the probability of a 1. This probability is related to the values of the explanatory variables $x_1, \ldots, x_k$.

We cannot, however, write this as $p = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ because the right-hand side is not constrained to lie in the interval $[0, 1]$, which it must if it is to represent a probability. One solution to this problem is to employ the *logit link function,* which is given by

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

and this leads to the equations

$$\frac{p}{1-p} = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}$$

and

$$p = \frac{\exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}}$$

for the *odds* $p/(1-p)$ and probability p, respectively. The right-hand side of the equation for p is now always between 0 and 1. Note that logistic regression is based on an ordinary regression relation between the logarithm of the odds in favor of the event occurring at a particular setting of the explanatory variables and the values of the explanatory variables $x_1, \ldots, x_k$. The quantity $\ln(p/(1-p))$ is referred to as the *log odds*.

The procedure for estimating the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ using this relation and carrying out tests of significance on these values is known as *logistic regression*. Typically, more sophisticated statistical methods than least squares are needed for fitting and inference in this context, and we rely on software such as Minitab to carry out the necessary computations.

In addition, other link functions are available in Minitab and are often used. In particular, the *probit link function* is given by

$$\Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

where Φ is the cumulative distribution function of the $N(0,1)$ distribution, and this leads to the relation

$$p = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

which is also always between 0 and 1. Choice of the link function can be made via a variety of goodness-of-fit tests available in Minitab, but we restrict our attention here to the logit link function.

16.2 Example

Suppose that we have the following 10 observations in columns C1–C3

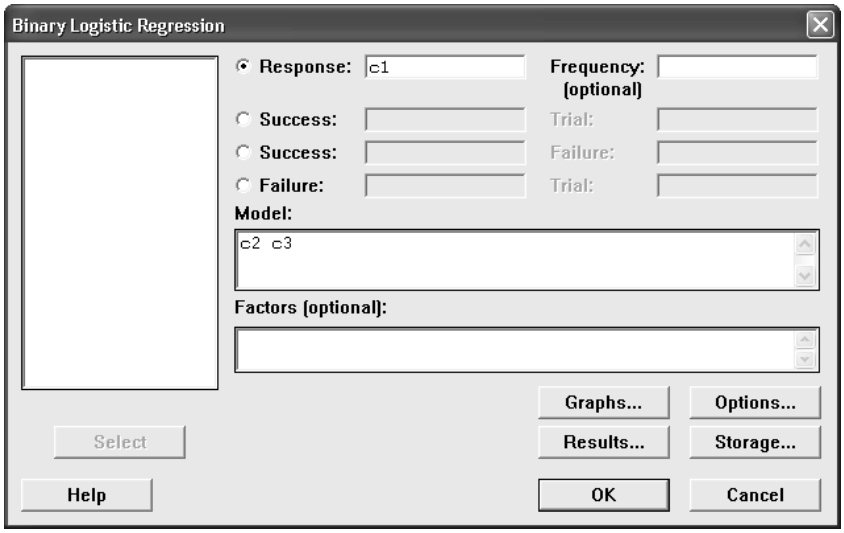
Row	C1	C2	C3
1	0	-0.65917	0.43450
2	0	0.69408	0.48175
3	1	-0.28772	0.08279
4	1	0.76911	0.59153
5	1	1.44037	2.07466
6	0	0.52674	0.27745
7	1	0.38593	0.14894
8	1	-0.00027	0.00000
9	0	1.15681	1.33822
10	1	0.60793	0.36958

where the response y is in C1, x_1 is in C2, and x_2 is in C3 and note that $x_2 = x_1^2$. We want to fit the model

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

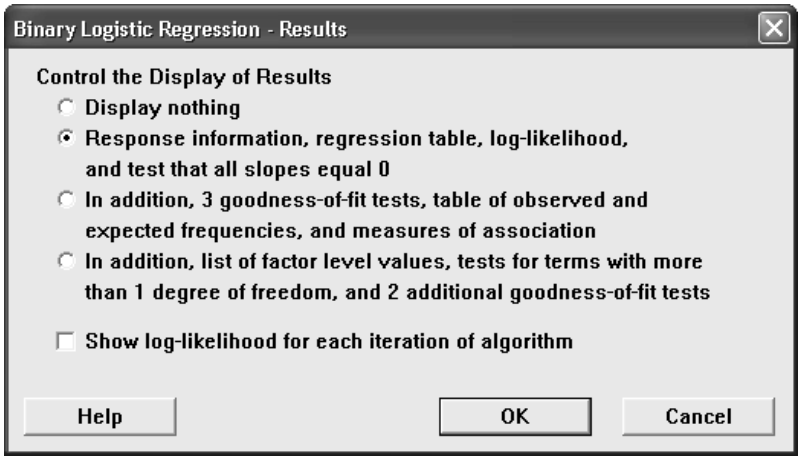
and conduct statistical inference concerning the parameters of the model.

Fitting and inference is carried out in Minitab using Stat ▶ Regression ▶ Binary Logistic Regression and filling in the dialog box as in Display 16.1.1.



Display 16.1.1: Dialog box for implementing a binary logistic regression.

Here, the Response box contains `c1` and the Model box contains C2 and C3. Clicking on the Results button brings up the dialog box in Display 16.1.2.



Display 16.1.2: The dialog box resulting from clicking on the 16.1.1.

We have filled in the radio button Response information, regression table, etc., as this controls the amount of output. The default output is more extensive and we chose to limit this. The following output is obtained:

```
Link Function:  Logit
Response Information

Variable Value Count
C1       1         6 (Event)
         0         4
        Total     10

Logistic Regression Table
                                                    Odds    95% CI
Predictor     Coef   StDev       Z      P     Ratio Lower Upper
Constant   0.522799 0.903137   0.58  0.563
C2         0.739955 1.60504    0.46  0.645    2.10  0.09 48.71
C3        -0.779614 1.58437   -0.49  0.623    0.46  0.02 10.23

Log-Likelihood = -6.598
Test that all slopes are zero:  G = 0.265, DF = 2,
                                              P-Value = 0.876
```

This gives estimates of the coefficients and their standard errors and the P-value for $H_0 : \beta_0 = 0$ versus $H_a : \beta_0 \neq 0$ as 0.563, the P-value for $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ as 0.645, and the P-value for $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$ as 0.623. Further, the test of $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ has P-value .876. In this example, there is no evidence of any nonzero coefficients. Note that $p = .5$ when $\beta_0 = \beta_1 = \beta_2 = 0$.

Also provided in the output is the estimate 2.10 for the odds ratio for x_1 (C2) and a 95% confidence interval $(.09, 48.71)$ for the true value. The odds ratio for x_1 is given by $\exp(\beta_1)$, which is the ratio of the odds at $x_1 + 1$ to the odds at x_1 when x_2 is held fixed or when $\beta_2 = 0$. Because there is evidence that $\beta_2 = 0$ (P-value $= .623$), the odds ratio has a direct interpretation here. Note, however, that if this wasn't the case the odds ratio would not have such an interpretation as it doesn't makes sense for x_2 to be held fixed when x_1 changes in this example as they are not independent variables. Similar comments apply to the estimate 0.46 for the odds ratio for x_2 (C3) and the 95% confidence interval $(.02, 10.23)$ for the true value of this quantity.

Many other aspects of fitting logistic regression models are available in Minitab and we refer the reader to Help for a discussion of these. Also available in Minitab are *ordinal logistic regression,* when the response takes more than two values and these are ordered, and *nominal logistic regression,* when the response takes more than two values and these are unordered. These can be accessed via Stat ▶ Regression ▶ Ordinal Logistic Regression and Stat ▶ Regression ▶ Nominal Logistic Regression, respectively.

16.3 Exercises

1. Generate a sample of 20 from the Bernoulli(.25) distribution. Pretending that we don't know p, compute a 95% confidence interval for this quantity.

Using this confidence interval, form 95% confidence intervals for the odds and the log odds.

2. Let x take the values -1, $-.5$, 0, .5, and 1. Plot the log odds

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

against x when $\beta_0 = 1$ and $\beta_1 = 2$. Plot the odds and the probability p against x.

3. Let x take the values -1, $-.5$, 0, .5, and 1. At each of these values, generate a sample of four values from the Bernoulli(p_x) distribution where

$$p_x = \frac{\exp\{1+2x\}}{1+\exp\{1+2x\}}$$

and let these values be the y response values. Carry out a logistic regression analysis of this data using the model.

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x$$

Compute a 95% confidence interval for β_1 and determine if it contains the true value. Similarly, form a 95% confidence interval for the odds ratio when x increases by 1 unit and determine if it contains the true value.

4. Let x take the values -1, $-.5$, 0, .5, and 1. At each of these values, generate a sample of four values from the Bernoulli(p_x) distribution where

$$p_x = \frac{\exp\{1+2x\}}{1+\exp\{1+2x\}}$$

and let these values be the y response values. Carry out a logistic regression analysis of this data using the model

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Test the null hypothesis $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$. Form a 95% confidence interval for the odds ratio for x. Does it make sense to make an inference about this quantity in this example? Why or why not?

5. Let x take the values -1, $-.5$, 0, .5, and 1. At each of these values, generate a sample of four values from the Bernoulli(.5) distribution. Carry out a logistic regression analysis of this data using the model

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Test the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$.

Chapter 17

Statistics for Quality: Control and Capability

New Minitab commands discussed in this chapter

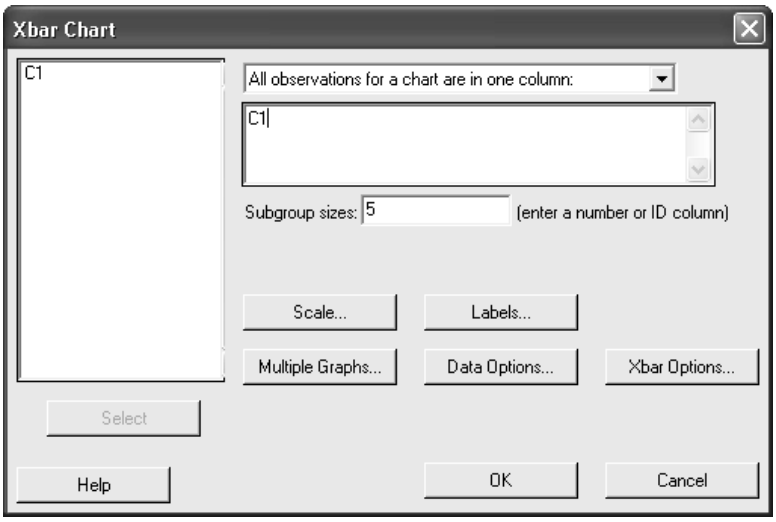
Stat ▶ Control Charts ▶ Attributes Charts ▶ P
Stat ▶ Control Charts ▶ Variables Charts for Subgroups ▶ S
Stat ▶ Control Charts ▶ Variables Charts for Subgroups ▶ Xbar

Control charts are used to monitor a process to ensure that it is under statistical control. There is a wide variety of such charts depending on the statistic used for the monitoring and the test used to detect when a process is out of control.

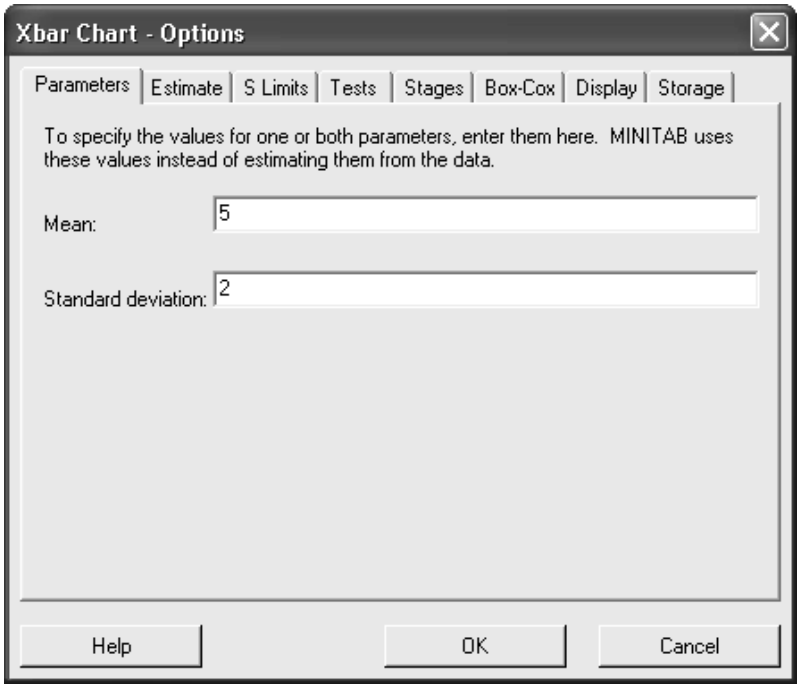
17.1 Producing $\bar{x}$ Charts

Suppose we have placed a random sample of 100 from the $N(5, 2)$ distribution in C1 and we want an $\bar{x}$ chart of this data. Then the command Stat ▶ Control Charts ▶ Variables Charts for Subgroups ▶ Xbar brings up the dialog box shown in Display 17.1.1. Here we have indicated that the data is in C1 and that we want the sample averages to be based on 5 observations (so there are 20 means). To control the placement of the LCL and UCL limits we clicked on Xbar Options ... to bring up the dialog box shown in Display 17.1.2. Here we asked that the center line be drawn at 5 and the standard deviation be set to 2 so that the LCL is $5 - 3(2/\sqrt{5}) = 2.3167$ and the UCL is $5 + 3(2/\sqrt{5}) = 7.6833$.

If we do not specify these values, then Minitab will estimate them from the data using the sample mean for the center line and the average of the sample standard deviations for the subgroups to determine the LCL and UCL. In particular, if $\bar{s}$ denotes the average standard deviation then the LCL equals $\bar{x} - 3\bar{s}/c_4$ and the UCL equals $\bar{x} + 3\bar{s}/c_4$, where c_4 is the constant defined in IPS that corrects for the bias in s, as an estimator of σ.

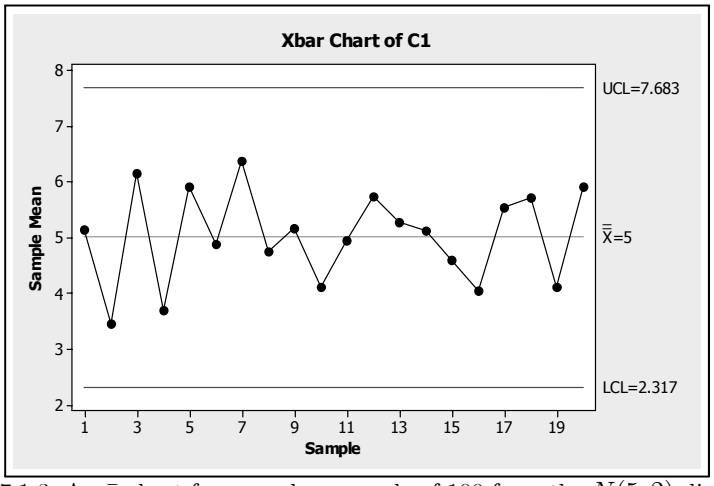


Display 17.1.1: Dialog box to create an $\bar{x}$ chart.



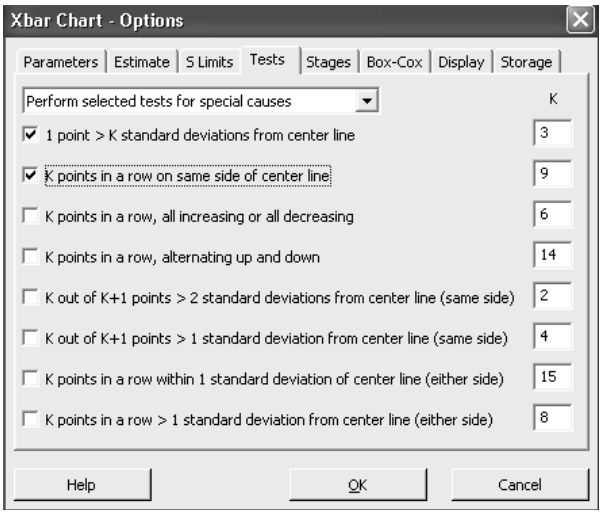
Display 17.1.2: Dialog box to control placement of center line and limits in an $\bar{x}$ chart.

Clicking on OK in both of these dialog boxes produces the $\bar{x}$ chart shown in Display 17.1.3. As expected, all the sample means lie within the limits.

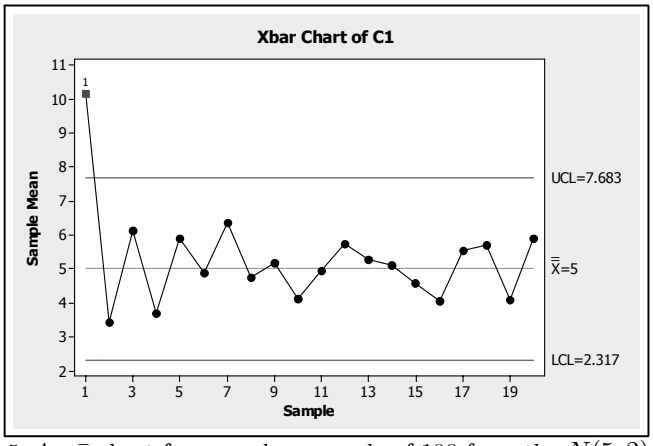


Display 17.1.3: An $\bar{x}$ chart for a random sample of 100 from the $N(5, 2)$ distribution.

We observe that the dialog box in Display 17.1.2 contains a tab labelled Tests. Clicking on this produces the dialog box shown in Display 17.1.4 where we have indicated that we want two tests to be carried out, namely, *1 point* $>$ *K standard deviations from center line* with $K = 3$ and *K points in a row on same side of center line* with $K = 9$. Clearly, the control chart shown in Display 17.1.3 passes both of these tests. Suppose, however, that we change the first sample observation to the value 30. Then using the dialog boxes shown in Displays 17.1.1, 17.1.2 and 17.1.4 produces the $\bar{x}$ chart shown in Display 17.1.5. Note that the first sample mean fails the first test and this is indicated on the chart by placing a 1 above that plotted mean. If any points had failed the second test, this would have been indicated by placing the number 2 above those plotted means, etc.



Display 17.1.4: Dialog box to choose tests to be performed in an $\bar{x}$ chart.



Display 17.1.5: An $\bar{x}$ chart for a random sample of 100 from the $N(5, 2)$ distribution where the first observation has been changed to be equal to 30.

The syntax of the corresponding session command **xbarchart** is

xbarchart E_1 E_2

where E_1 is a column containing the data and E_2 is either a constant, indicating how many observations are used to define a subgroup, or a column of values, indicating how the elements of E_1 are to be grouped for the calculation of the means. Minitab then produces the center line and control limits based on the data in E_1. When E_2 equals 1, σ cannot be estimated using standard deviations and an alternative estimator is used.

There are various subcommands that can be used with **xbarchart**. In particular, we can provide **mu** and **sigma** to specify the population mean and standard deviation. For example, the commands

```
MTB > xbarchart c1 5;
SUBC> mu 5;
SUBC> sigma 2.
```

produce the chart shown in Display 17.1.3.

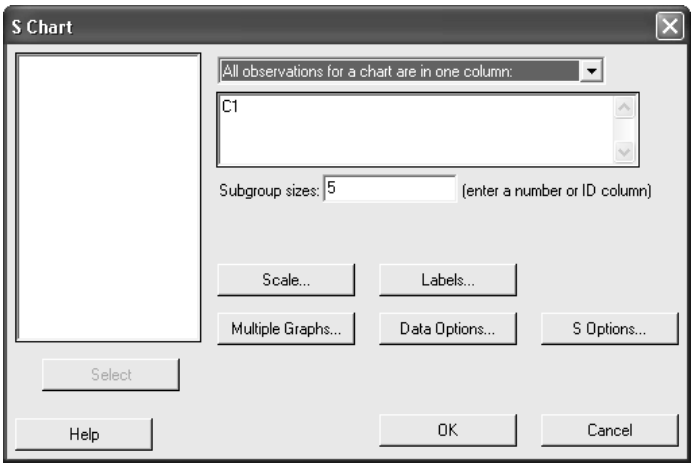
Using the **test** subcommand, various tests for control can be carried out. For example,

```
MTB > xbar c1 5;
SUBC> test 1.
```

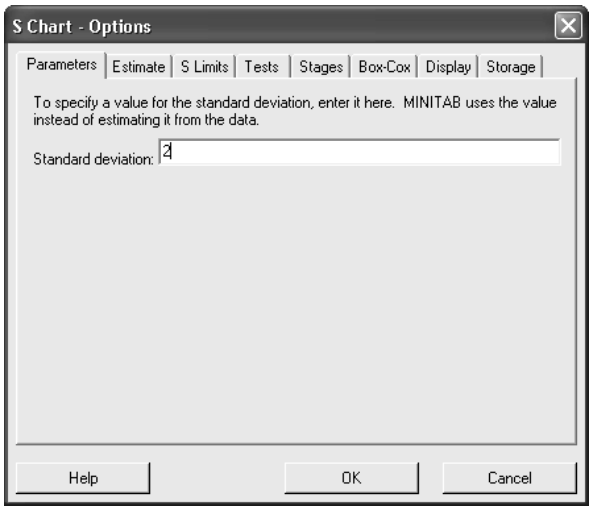
breaks the data into subgroups of size 5 and checks to see if any of the points are outside the control limits. The subcommand `test 2` checks to see if there are 9 points in a row on the same side of the center line, `test 3` checks to see if there are 6 points in a row all increasing or all decreasing. There are a total of 8 tests like this, all looking for patterns. The subcommand `test 1:8` performs all 8 tests.

17.2 Producing S Charts

Suppose we have placed a random sample of 100 from the $N(5, 2)$ distribution in C1 and we want an S chart of this data. Then the command Stat ▶ Control Charts ▶ Variables Charts for Subgroups ▶ S brings up the dialog box shown in Display 17.2.1. Here we have indicated that the data is in C1 and that we want the sample standard deviations to be based on 5 observations (so there are 20 standard deviations). To control the placement of the LCL and UCL limits we clicked on S Options ... to bring up the dialog box shown in Display 17.2.2. Here we set $\sigma = 2$ so that the center line and the LCL and UCL limits are determined by this. If we don't specify the value for σ, then this parameter is estimated from the data.

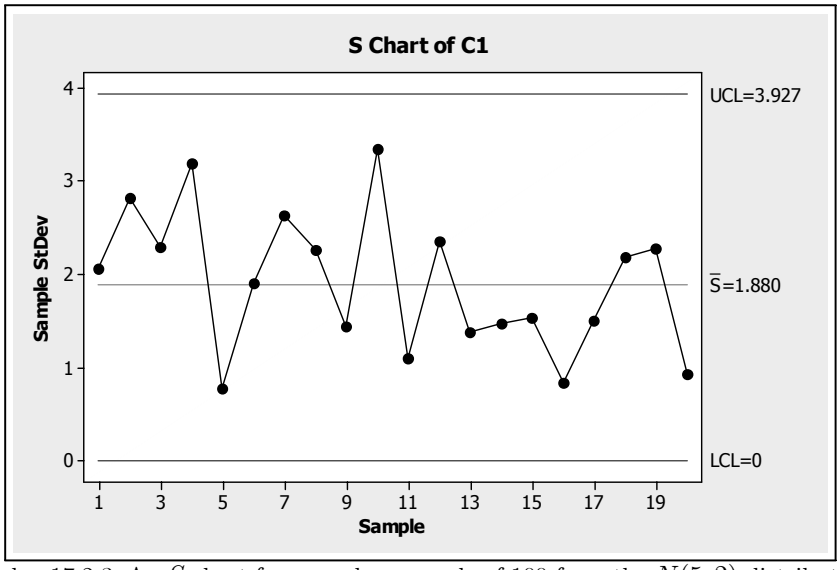


Display 17.2.1: Dialog box to create an S chart.



Display 17.2.2: Dialog box to control center line and limits in an S chart.

Clicking on OK in both of these dialog boxes produces the S chart shown in Display 17.2.3. As expected, all the standard deviations lie within the limits.



Display 17.2.3: An S chart for a random sample of 100 from the $N(5, 2)$ distribution.

We observe that the dialog box in Display 17.2.2 contains a tab labelled Tests. As with $\bar{x}$ charts (Display 17.1.4) we can select several tests to be performed to assess whether or not the process is in control.

The syntax of the corresponding session command **schart** is

schart E_1 E_2

where E_1 is a column containing the data and E_2 is either a constant, indicating how many observations are used to define a subgroup, or a column of values, indicating how the elements of E_1 are to be grouped for the calculation of the standard deviations. Minitab then produces the center line and control limits based on the data in E_1. When E_2 equals 1, σ cannot be estimated using standard deviations and an alternative estimator is used. There are various subcommands that can be used with **schart**. For example, the commands

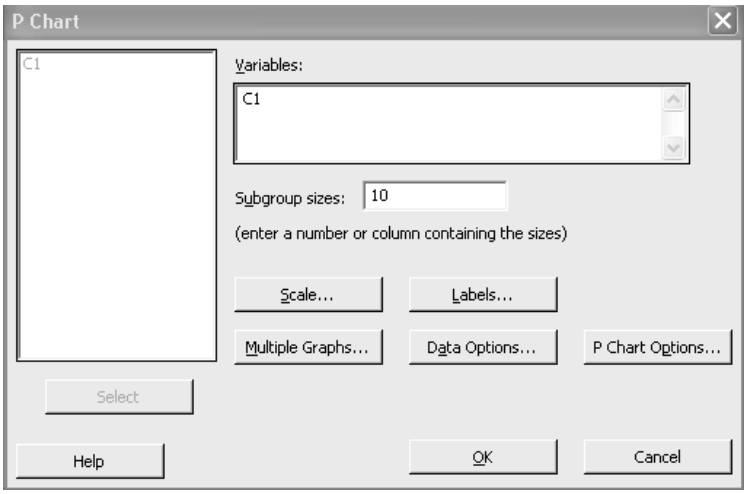
```
MTB > schart c1 5;
SUBC> sigma 2.
```

produces the control chart of Display 17.2.3.

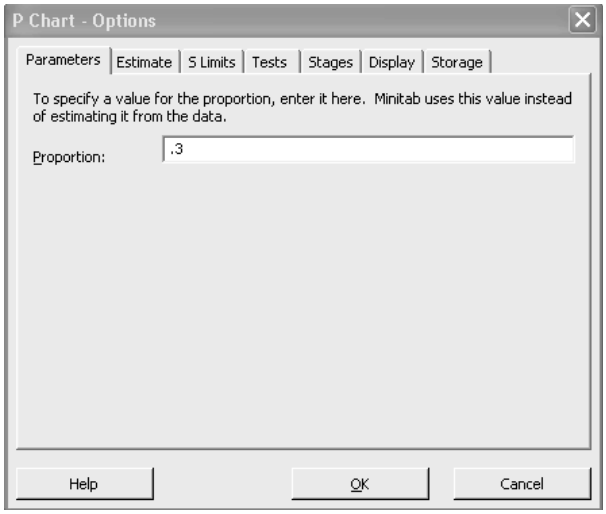
17.3 Producing p Charts

A p chart is appropriate when a response is coming from a Binomial(n, p) distribution; for example, the count of the number of defectives in a batch of size n, and we use the proportion of defectives $\hat{p}$ to control the process. Suppose

we have placed a random sample of 50 from the Binomial$(10, .3)$ distribution in C1 and we want a p chart of this data. Then the command Stat ▶ Control Charts ▶ Attributes Charts ▶ P brings up the dialog box shown in Display 17.3.1. Here we have indicated that the data is in C1 and that these counts are based on 10 observations. To control the placement of the LCL and UCL limits we clicked on P Chart Options to bring up the dialog box shown in Display 17.3.2. Here we asked that limits be determined by setting $p = .3$ so that the center line is at .3, the LCL is $\max\left\{.3 - 3\sqrt{.3(.7)/10}, 0\right\} = 0.0$, and the UCL is $.3 + 3\sqrt{.3(.7)/10} = 0.73474$. If we don't specify the value for p then this parameter is estimated from the data and the center line and limits depend on the data.

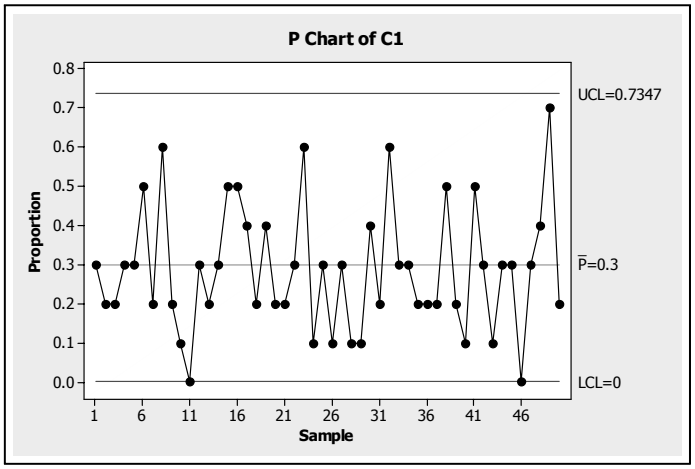


Display 17.3.1: Dialog box to create a p chart.



Display 17.3.2: Dialog box to control center line and limits in a p chart.

Clicking on OK in these dialog boxes produces the p chart shown in Display 17.3.3. We see from this that the process seems to be in control as we might expect.



Display 17.3.3: A p chart for a random sample of 50 from the Binomial$(10, .3)$ distribution.

The syntax of the corresponding session command **pchart** is

pchart E_1 E_2

where E_1 is a column containing the data and E_2 is a constant, indicating how many observations the counts are based on. Minitab then produces the center line and control limits based on the data in E_1. There are various subcommands that can be used with **pchart**. For example, the commands

```
MTB > pchart C1 10;
SUBC> P .3.
```

produce the plot shown in Display 17.3.3.

17.4 Exercises

1. Generate a sample of 100 from a Student(1). Make an $\bar{x}$ chart for this data based on subgroups of size 5 with $\mu = 0$ and $\sigma = 1$. What tests for control are failed?

2. For the data in Exercise 1, make an $\bar{x}$ chart based on subgroups of size 5 using estimates of μ and σ. What tests for control are failed?

3. For the data in Exercise 1, make an S chart based on subgroups of size 5 using $\sigma = 1$. What tests for control are failed?

4. For the data in Exercise 1, make an S chart based on subgroups of size 5 using an estimate of σ. What tests for control are failed?

5. Generate a sample of 100 from a Binomial$(15, .1)$ distribution. Make a p chart for this data. What tests for control are failed?

6. Generate a sample of 50 from a Binomial$(15, .1)$ distribution followed by a sample of 50 from a Binomial$(15, .8)$ distribution. Make a p chart for this data. What tests for control are failed?

Chapter 18

Time Series Forecasting

New Minitab commands discussed in this chapter

Graph ▶ Time Series Plot
Stat ▶ Time Series ▶ ARIMA
Stat ▶ Time Series ▶ Decomposition
Stat ▶ Time Series ▶ Single Exp Smoothing
Stat ▶ Time Series ▶ Lag
Stat ▶ Time Series ▶ Moving Average
Stat ▶ Time Series ▶ Trend Analysis

18.1 Time Series Plots

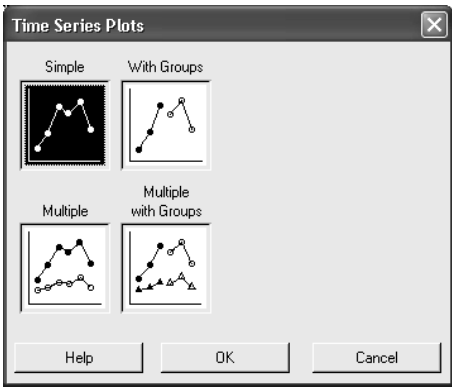
Often, data are collected sequentially in time. In such a context, it is instructive to plot the values of quantitative variables against time in a time series plot. For this we use the Graph ▶ Time Series Plot command.

Suppose that we obtain the following series of 50 successive daily prices recorded for a commodity where the time proceeds along rows. These data values are placed in C1 and are used throughout this chapter.

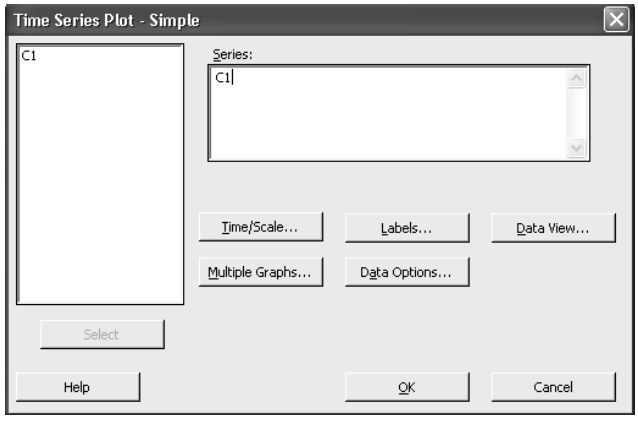
39	35	39	38	37	38	42	41	41	42
46	47	51	54	54	53	57	49	46	43
51	43	51	45	34	36	36	37	34	32
28	31	28	27	29	28	20	18	22	23
29	29	30	25	21	27	28	29	33	34

The Graph ▶ Time Series Plot command brings up the dialog box shown in Display 18.1.1. Clicking on Simple and OK brings up the dialog box shown in Display 18.1.21 where we have asked for a time series plot of the variable C1. This produces the time plot shown in Display 18.1.3 where price is plotted against day. There are various options available to modify the presentation of this graph.

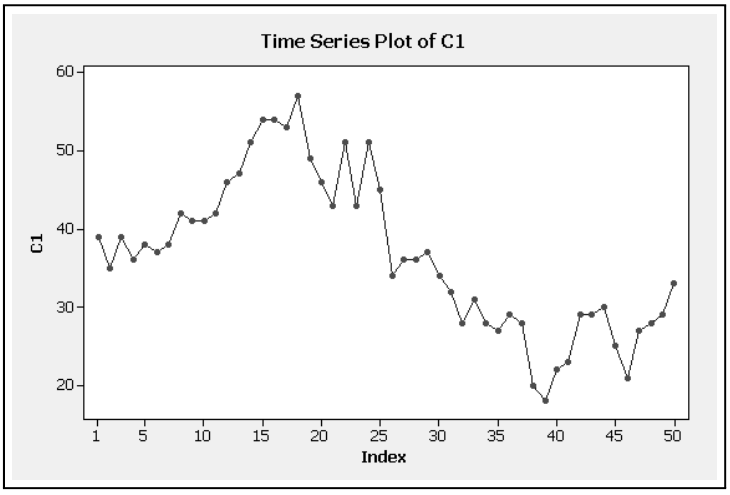
There is also a corresponding session command **tsplot.** We refer the reader to **help** for more discussion of this.



Display 18.1.1: First dialog box for producing a time series plot.



Display 18.1.2: Dialog box for a time series plot of the variable C1.



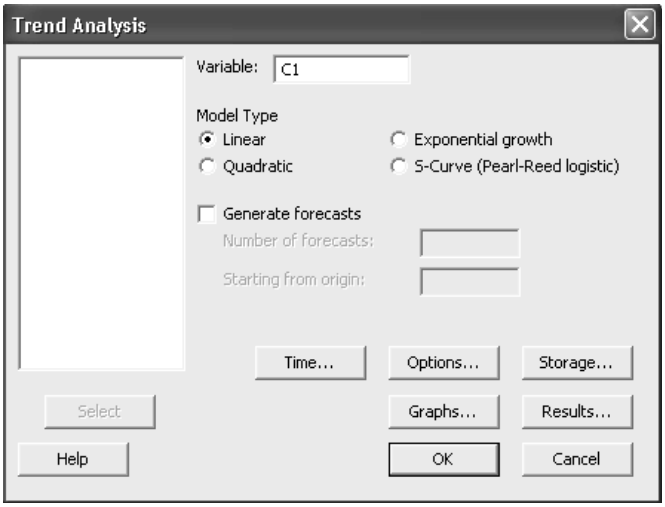
Display 18.1.3: Time series plot of the variable `C1`.

18.2 Trend Analysis

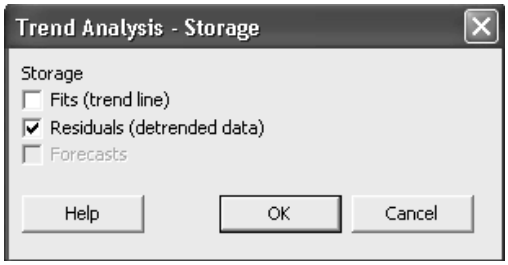
The command Stat ▶ Time Series ▶ Trend Analysis can be used to fit curves to a time series to determine the trend. This command brings up the dialog box in Display 18.2.1, where we have asked for a linear trend analysis for the variable C1. Clicking on the Storage button brings up the dialog box in Display 18.2.2 where we have asked for the residuals to be stored. These choices produce the output

```
Trend Analysis for C1
Data C1
Length 50
NMissing 0
Fitted Trend Equation
Yt = 48.17 - 0.452245*t
```

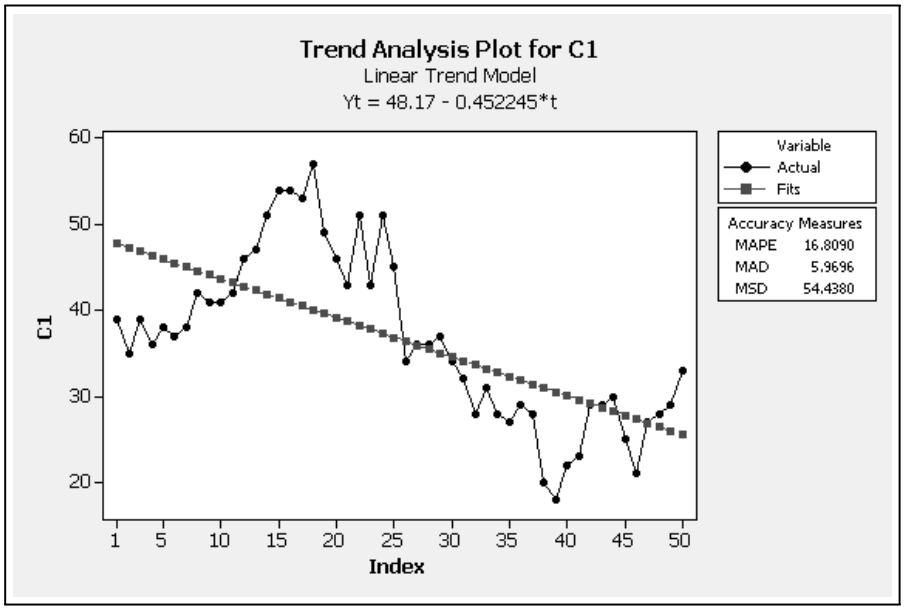
in the session window and the plot in Display 18.2.3. Also the residuals are stored in the variable `RESI1`. The fitted line is given by $y_t = 48.17 - 0.452245t$, which indicates a trend of decreasing prices with time.



Display 18.2.1: Dialog box for a trend analysis of the variable C1.



Display 18.2.2: Dialog box for selected items to be stored in a trend analysis.



Display 18.2.3: Plot of fitted linear trend and the time series C1.

Note that the residuals can be plotted in a time series plot to check for autocorrelation and, in this case, indicates a clear autocorrelation. Also, we can use the Stat ▶ Time Series ▶ Lag command to place the lagged residuals in another column and then graph the residuals against the lagged residuals in a scatterplot as another check for autocorrelation.

There is also a corresponding session command **trend.** We refer the reader to **help** for more discussion of this command.

18.3 Seasonality

Suppose that the data represents 10 weeks of 5 successive trading days and we want to see if there is any evidence of a weekly pattern to the pricing of this commodity. The command Stat ▶ Time Series ▶ Decomposition brings up the dialog box of Display 18.3.1. We have selected to fit an additive model, trend plus seasonality, as opposed to a multiplicative model, trend times seasonality, by filling in Additive under Model Type and putting 5 in the Seasonal length box. Clicking on OK produces the output

```
Time Series Decomposition for C1
Additive Model
Data C1
Length 50
NMissing 0
Fitted Trend Equation
Yt = 48.20 - 0.453397*t
```

```
Seasonal Indices
Period     Index
     1     -1.26
     2      0.54
     3      1.04
     4      0.14
     5     -0.46
```

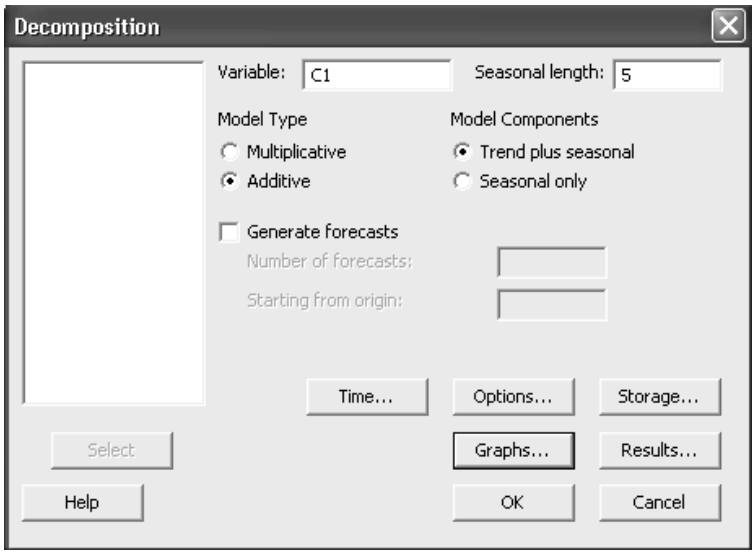
in the session window. This gives the trend equation $y_t = 48.20 - 0.453397t$ and the estimates of the seasonal effects. So the fitted value on day 32 is $48.20 - (0.453397)13 + 1.04 = 43.346$. Note that we could also have generated forecasts by clicking in the Generate forecasts box and entering a number in the Number of forecasts box. Further, various plots are provided. In Display 18.3.2 we have a plot of the original data, the trend line, and the trend plus seasonal curve. From this we can see that there is little if any benefit of including the seasonal term and so we have evidence against such a seasonal effect existing.

Note that we can also fit a multiplicative model by filling in the Multiplicative button under Model Type in the dialog box of Display 18.3.1. There are also a number of options for storing various quantities, plots of residuals, etc.

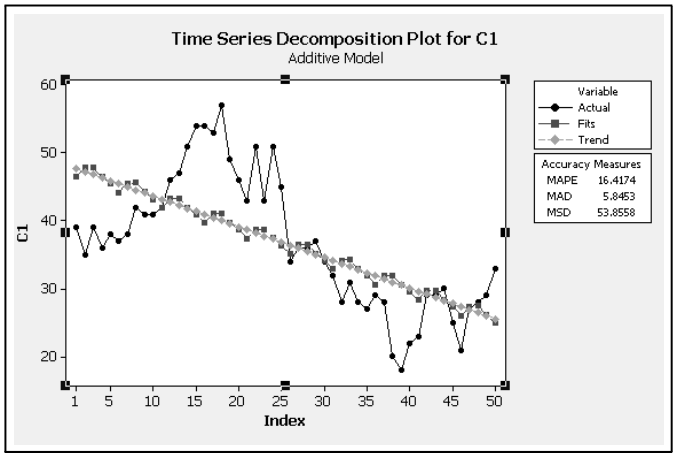
There is also a corresponding session command **decomposition.** For example the command

```
MTB > decomposition c1 5
```

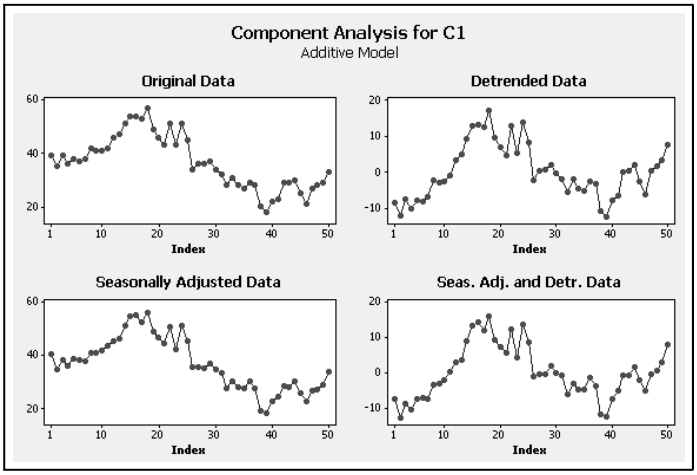
fits the multiplicative model with seasonal length 5. We refer the reader to **help** for more discussion of this command.



Display 18.3.1: Dialog box for fitting an additive trend and seasonality model.



Display 18.3.2: Plot of additive trend and seasonality model for data in C1.



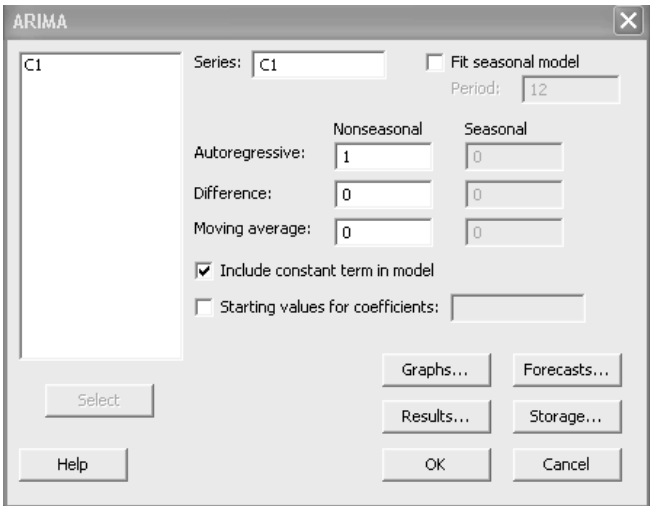
Display 18.3.3: Plot of original data, detrended data, seasonally adjusted data, and seasonally adjusted and detrended data for additive trend and seasonality model for data in C1.

18.4 Autoregressive Model

To fit the first-order autoregressive model $y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$ we use the Stat ▶ Time _Series ▶ ARIMA command with the dialog box as in Display 18.4.1. We have requested that an AR(1) model be fitted by placing a 1 in the Nonseasonal Autoregressive box and 0's in the Difference and Moving Average boxes. This produced the output

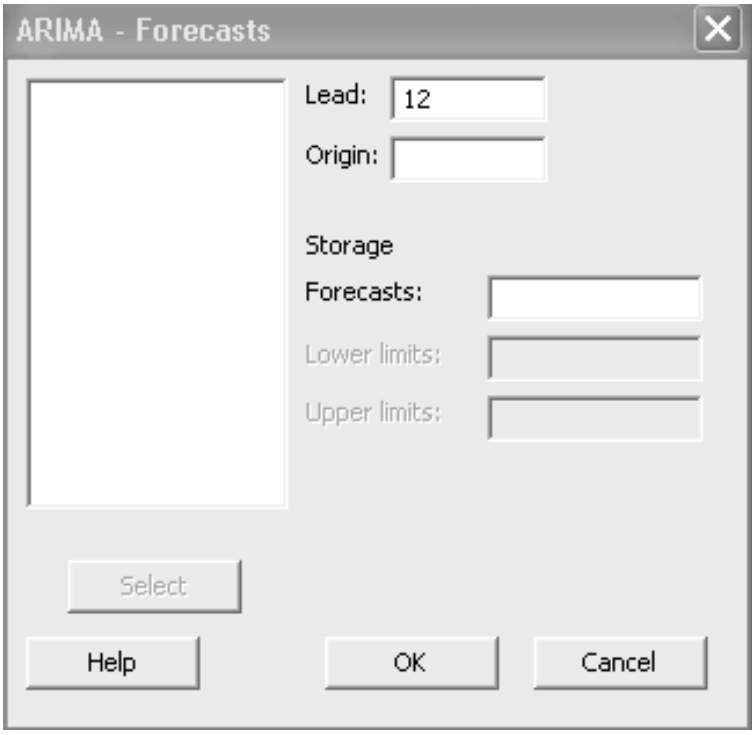
```
ARIMA Model:  C1
Estimates at each iteration
Iteration      SSE    Parameters
        0  4014.85  0.100 33.066
        1  2940.25  0.250 27.543
        2  2083.17  0.400 22.021
        3  1443.62  0.550 16.501
        4  1021.58  0.700 10.984
        5   817.04  0.850  5.475
        6   796.02  0.912  3.188
        7   795.94  0.916  3.067
        8   795.94  0.916  3.059
Relative change in each estimate less than 0.0010

Final Estimates of Parameters
    Type     Coef    SE Coef      T      P
AR 1       0.9161     0.0585  15.66  0.000
Constant   3.0591     0.5759   5.31  0.000
Mean      36.443      6.860
Number of observations:  50
```

in the session window together with some additional output connected with testing various hypotheses concerning the model. The first table shows the results of the iterative fitting algorithm for computing the estimates of β_0 and β_1, which leads to the final estimates $\hat{y}_t = 3.0591 + 0.916\hat{y}_{t-1}$. If we click on the Storage button, in the dialog box of Display 18.4.1, then we see that we can store the fitted values; plotting the fitted values and the original series on the same plot show that this model gives a reasonable fit.



Display 18.4.1: Dialog box for fitting an AR(1) model to the data in C1..

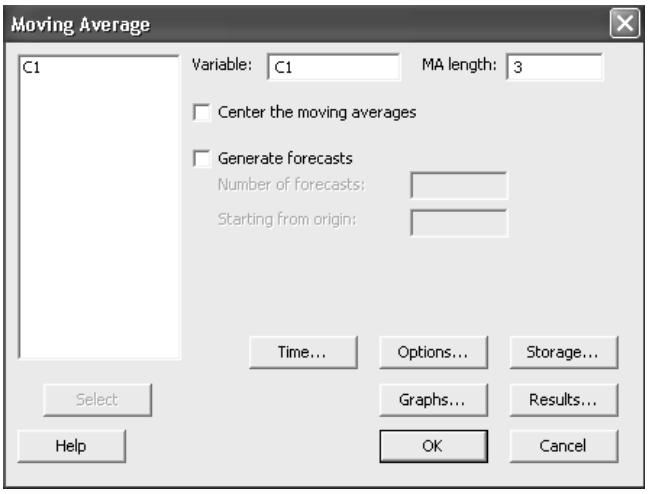
If we wish to obtain Forecasts based on the fitted model, then we click on the Forecasts button and fill in the dialog box as in Display 18.4.2 and put the number of forecasts in the Lead box. We have asked for 12 forecasts for trading days 51 through 62. These forecasts are printed in the session window.



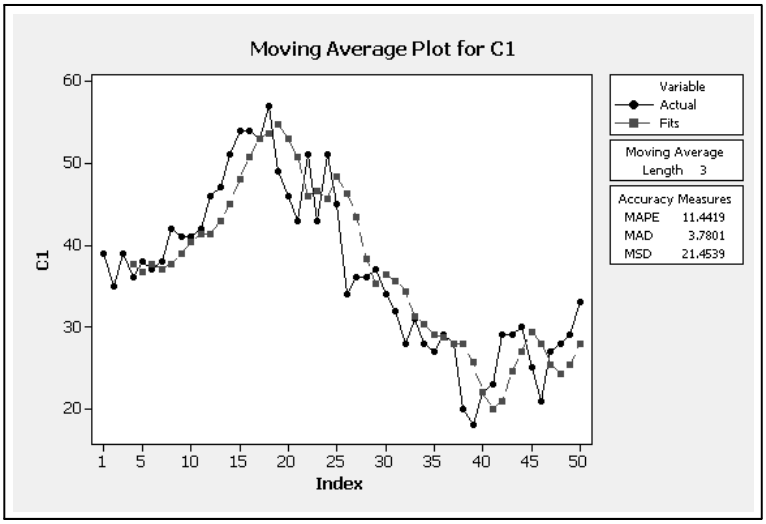
Display 18.4.2: Dialog box for selecting the number of forecasts when fitting an AR(1) model. .

18.5 Moving Averages

To fit a moving average to the series we use the command Stat ▶ Time Series ▶ Moving Average with the dialog box as in Display 18.5.1. Here we have asked to compute a series of moving averages based on an average of 5 values. This produces the graph in Display 18.5.2 where the original series and the fitted values (moving average of immediately preceding values) are plotted.



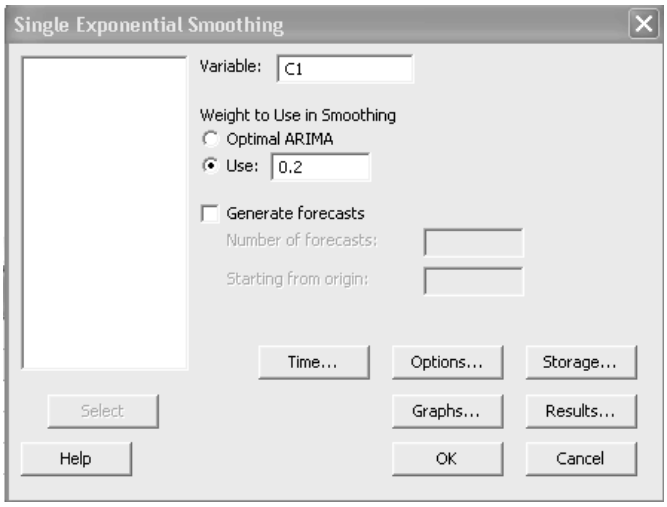
Display 18.5.1: Dialog box for fitting a moving average of span 5 to the data in C1.



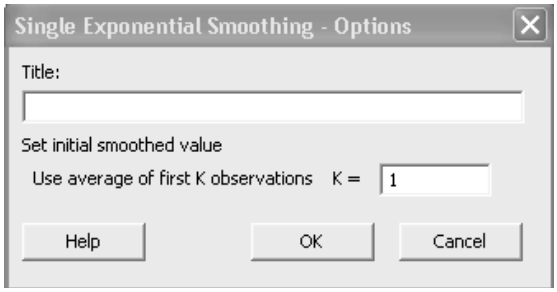
Display 18.5.2: Plot of series and moving average of span 3 for data in C1..

18.6 Exponential Smoothing

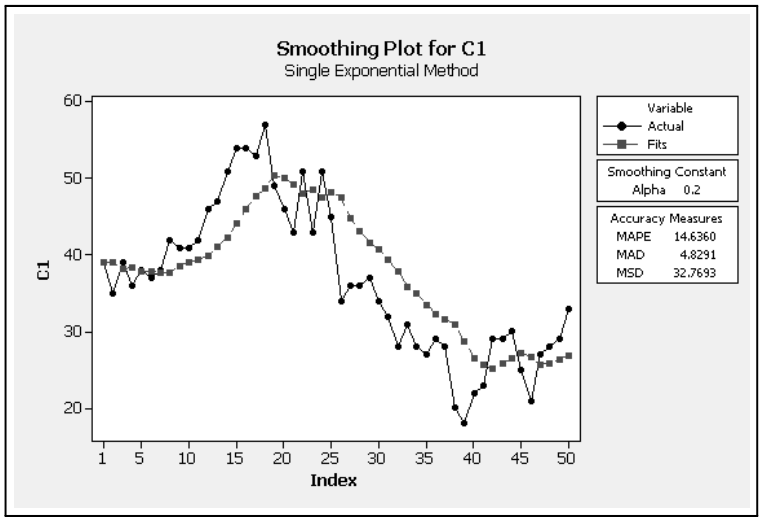
For exponential smoothing we use the Stat ▶ Time _Series ▶ Single Exp Smoothing command and the dialog box as in Display 18.6.1 where we have selected the weight $w = .2$ for the smoothing. Clicking on the Options button brings up the dialog box in Display 18.6.2 where we have requested that the initial value be y_1. If we set k equal to a larger integer we will take the initial value to the average of the first k observations.



Display 18.6.1: Dialog box for exponential smoothing the data in C1 with $w = .2$.



Display 18.6.2: Dialog box to set initial value for exponential smoothing.



Display 18.6.3: Plot of original series and exponentially smoothed series, with $w = .2$, for the data in C1.

18.7 Exercises

1. For the data in section 18.1, produce a time series plot of the residuals obtained from a linear trend analysis and comment on the existence of autocorrelation.

2. For the data in section 18.1, produce a scatterplot of successive residuals obtained from a linear trend analysis and comment on the existence of autocorrelation.

3. For the data in section 18.1, fit a multiplicative model for trend and seasonality. Compare the results with the fit of an additive model.

4. For the data in section 18.1, plot the fitted values for an AR(1) model together with the original series in a time series plot. Comment on how well the model fits the series.

5. For the data in section 18.1, plot moving average series of spans 5 and 7 and compare these to the plot in Display 18.5.2. Obtain a forecast for the next trading day.

6. For the data in section 18.1, plot the exponentially smoothed series for $w = .01, .4, .6, .8., .9$, and .99. What do you observe about the smoothed series?

Appendix A

Projects

The basic structural component of Minitab is the worksheet. When working on a project, it may make sense to have your data in several worksheets so that similar variables are grouped together. Also, you may wish to save plots associated with the worksheets so that everything can be obtained via a single reference. Worksheets and graphs can be grouped together into *projects.* Projects are given names and are stored in a file with the supplied name and the file extension `.mpj`.

To open a new project use File ▶ New and choose Minitab Project and click OK. If you want to open a previously saved project, use File ▶ Open Project and choose the relevant project from the list. To save a project use File ▶ Save Project if the project already has a name (or you wish to use the default of `minitab`) or File ▶ Save Project As if you wish to give the project a name. Not only are the contents of all worksheets and graphs saved, but the contents of the History folder in the Project window are saved as well and are available when the project is reopened. You can also supply a description of the project using File ▶ Project Description and filling in the dialog box. Note that a description of a worksheet can also be saved using Editor ▶ Worksheet ▶ Description. When you attempt to open a new project or exit Minitab, you will be asked if you wish to save the contents of the current project.

Now suppose that in the project `evans` we have a single worksheet containing 100 numeric values in each of C1 and C2 and have produced a scatterplot of C2 against C1. We open a new worksheet using File ▶ New and choose Minitab Worksheet and click OK. There are now two worksheets associated with the project called Worksheet1 and Worksheet2. Suppose that we also place 100 numeric values in C1 and C2 in Worksheet 2 and again plot C2 against C1. We then have two plots associated with the project `evans` called Worksheet 1: Plot C2*C1 and Worksheet 2: Plot C2*C1. These will all appear as individual windows on your screen, perhaps with some hidden, and any one in particular can be made active by clicking in that window or by clicking on the relevant entry in the list obtained when you use Window. You can also save individual worksheets in the project to files outside the project when a particular worksheet

is active using File ▶ Save Current Worksheet As. Similarly, when a graph window is active, a graph in the project can be saved to a file outside the project using File ▶ Save Graph As.

With multiple worksheets in a project, it is easy to move data between worksheets using cut, copy, and paste operations. For example, suppose that we want to copy C1 and C2 of Worksheet 1 into C3 and C4 of Worksheet 2. With Worksheet 1 active, highlight the entries in C1 and C2, use Edit ▶ Copy Cells, make Worksheet 2 active, click in the first cell of C3, and use Edit ▶ Paste Cells.

It is possible to see what a project contains without opening it. To do this use File ▶ Open Project, click on the project to be previewed, and click on the Preview button. Similarly, worksheets can be previewed using File ▶ Open Worksheet, clicking on the worksheet to be previewed, and clicking on the Preview button.

Appendix B

Functions in Minitab

B.1 Mathematical Functions

Here is a list and description of some of the mathematical and statistical functions available in Minitab. All of these functions operate on each element of a column and return a column of the same length. Let $(x_1, \ldots, x_n)$ denote a column of length n. These functions can be applied only to numerical variables.

abs - Computes the absolute value, $(|x_1|, \ldots, |x_n|)$.
antilog - Computes the inverse of the base 10 logarithm, $(10^{x_1}, \ldots, 10^{x_n})$.
acos - Computes the inverse cosine function, $(\arccos(x_1), \ldots, \arccos(x_n))$.
asin - Computes the inverse sine function, $(\arcsin(x_1), \ldots, \arcsin(x_n))$.
atan - Computes the inverse tangent function, $(\arctan(x_1), \ldots, \arctan(x_n))$.
cos - Computes the cosine function when angle is given in radians, $(\cos(x_1), \ldots, \cos(x_n))$.
ceiling - Computes the smallest integer bigger than a number, $(\lceil x_1 \rceil, \ldots, \lceil x_n \rceil)$.
degrees - Computes the degree measurement of an angle given in radians.
exp - Computes the exponential function, $(e^{x_1}, \ldots e^{x_n})$.
floor - Computes the greatest integer smaller than a number, $(\lfloor x_1 \rfloor, \ldots, \lfloor x_n \rfloor)$.
gamma - Computes the gamma function, $(\Gamma(x_1), \ldots, \Gamma(x_n))$; note that for nonnegative integer x, $\Gamma(x+1) = x!$.
lag - Computes the column $(*, x_1, \ldots, x_{n-1})$.
ln - Computes the natural logarithm function, $(\ln(x_1), \ldots, \ln(x_n))$.
lngamma - Computes the log-gamma function, $(\ln\Gamma(x_1), \ldots, \ln\Gamma(x_n))$; note that for nonnegative integer x, $\ln\Gamma(x+1) = \sum_{i=1}^{x} \ln(i)$.
logten - Computes the base 10 logarithm function, $(\log_{10}(x_1), \ldots, \log_{10}(x_n))$.
nscore - Computes the normal scores function; see **help**.
parproducts - Computes the column of partial products, $(x_1, x_1x_2, \ldots, x_1 \cdot \cdots \cdot x_n)$.

parsums - Computes the column of partial sums,
$(x_1, x_1 + x_2, \ldots, x_1 + \cdots x_n)$.
radians - Computes the radian measurement of an angle given in degrees.
rank - Computes the ranks of the column entries, $(r_1, \ldots r_n)$.
round - Computes the nearest integer function $i(x)$ with rounding up at .5,
$(i(x_1), \ldots, i(x_n))$; see **help** for more details on this function.
signs - Computes the sign function

$$s(x) = \begin{cases} -1 & if \quad x < 0 \\ 0 & if \quad x = 0 \\ 1 & if \quad x > 0 \end{cases}$$

$(s(x_1), \ldots, s(x_n))$.
sin - Computes the sine function when the angle is given in radians,
$(\sin(x_1), \ldots, \sin(x_n))$.
sort - Computes the column consisting of the sorted (ascending) column entries,
$\left(x_{(1)}, \ldots, x_{(n)}\right)$.
sqrt - Computes the square root function, $\left(\sqrt{x_1}, \ldots, \sqrt{x_n}\right)$.
tan - Computes the tangent function when the angle is given in radians,
$(\tan(x_1), \ldots, \tan(x_n))$.

B.2 Column Statistics

Let $(x_1, \ldots, x_n)$ denote a column of length n. Output is written on the screen or in the Session window and can be assigned to a constant. The general syntax for column statistic commands is

column statistic name(E_1)

where the operation is carried out on the entries in column E_1 and output is written to the screen unless it is assigned to a constant using the **let** command.

max - Computes the maximum of a column, $x_{(n)}$.
mean - Computes the mean of a column, $\bar{x} = (x_1 + \cdots x_n)/n$.
median - Computes the median of a column (see Chapter 1).
min - Computes the minimum of a column, $x_{(1)}$.
n - Computes the number of nonmissing values in the column.
nmiss - Computes the number of missing values in the column.
range - Computes the difference between the smallest and largest value in a column,
$x_{(n)} - x_{(1)}$.
ssq - Computes the sum of squares of a column, $x_1^2 + \cdots + x_n^2$.
stdev - Computes the standard deviation of a column,

$$s = \sqrt{\frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right]}.$$

sum - Computes the sum of the column entries, $x_1 + \cdots x_n$.

B.3 Row Statistics

Let $(x_1, \ldots, x_n)$ denote a row of length n. The general syntax is

row statistic name $\mathrm{E}_1 \ldots \mathrm{E}_m$ E_{m+1}

where the operations are carried out on the rows in columns $\mathrm{E}_1, \ldots, \mathrm{E}_m$ and the output is placed in column E_{m+1}.

rmax - Computes the maximum of a row, $x_{(n)}$.
rmean - Computes the mean of a row, $\bar{x} = (x_1 + \cdots x_n)/n$.
rmiss - Computes the number of missing values in the row.
rn - Computes the number of nonmissing values in the row.
rrange - Computes the difference between the smallest and largest value in a row,

$x_{(n)} - x_{(1)}$.

rssq - Computes the sum of squares of a row, $x_1^2 + \cdots + x_n^2$.
rstdev - Computes the standard deviation of a row,

$$s = \sqrt{\frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right]}.$$

rsum - Computes the sum of the row entries, $x_1 + \cdots x_n$.

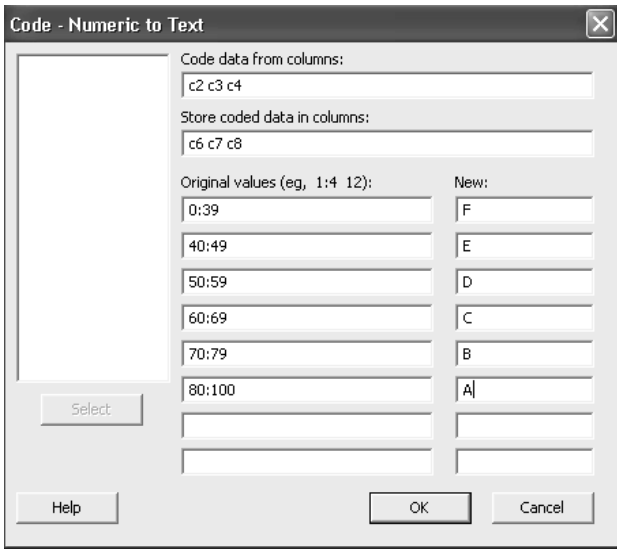
Appendix C

More Minitab Commands

In this section, we discuss some commands that can be very helpful in certain applications. We will make reference to these commands at appropriate places throughout the manual. It is probably best to wait to read these descriptions until such a context arises.

C.1 Coding

The Data ▶ Code command is used to recode columns. By this we mean that data entries in columns are replaced by new values according to a coding scheme that we must specify. You can recode numeric into numeric, numeric into text, text into numeric, or text into text by choosing an appropriate subcommand. For example, suppose in the `marks` worksheet (Display I.4) we want to recode the grades in C2, C3, and C4 so that any mark in the range 0–39 becomes an F, every mark in the range 40–49 becomes an E, every mark in the range 50–59 becomes a D, every mark in the range 60–69 becomes a C, every mark in the range 70–79 becomes a B, every mark in the range 80–100 becomes an A, and the results are placed in columns C6, C7, and C8, respectively. Then the command Data ▶ Code ▶ Numeric to Text brings up the dialog box shown in Display C.1.1. The ranges for the numeric values to be recoded to a common text value are typed in the Original values box, and the new values are typed in the New box. Note that we have used a shorthand for describing a range of data values. Because the sixth entry of C4 is *—i.e., it is missing—this value is simply recoded as a blank. You can also recode missing values by including * in one of the Original values boxes. If a value in a column is not covered by one of the values in the Original values boxes, then it is simply left the same in the new column.



Display C.1.1: Dialog box for recoding numeric values to text values.

Note that this menu command restricts the number of new code values to 8. The session command **code** allows up to 50 new codes. For example, suppose in the `marks` worksheet we want to recode the grades in C2, C3, and C4 so that any mark in the range 0–9 becomes a 0, every mark in the range 10–19 becomes 10, etc., and the results are placed in columns C6, C7, and C8. The following command

```
MTB >code(0:9) to 0 (10:19) to 10 (20:29) to 20 (30:39) to 30 &
CONT>(40:49) to 40 (50:59) to 50 (60:69) to 60 (70:79) to 70 &
CONT>(80:89) to 80 (90:99) to 90 for C2-C4 put in C6-C8
```

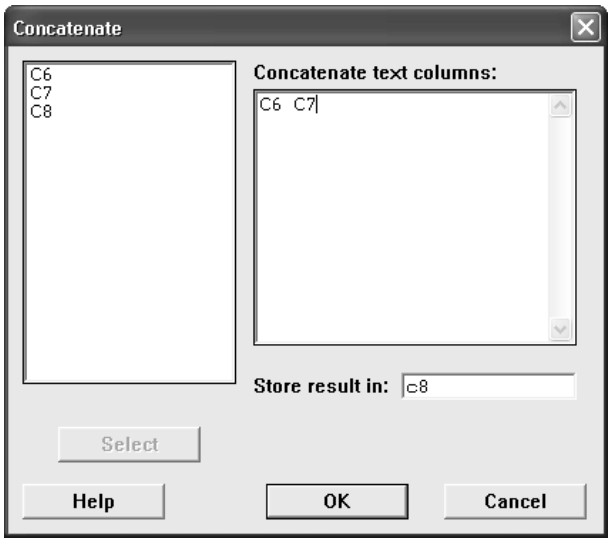
accomplishes this. Note the use of the continuation symbol &, as this is a long command. The general syntax for the **code** command is

code (V_1) to $code_1$ ... (V_n) to $code_n$ for E_1 ... E_m put in E_{m+1} ... E_{2m}

where V_i denotes a set of possible values and ranges for the values in columns E_1 ... E_m that are all coded as the number $code_i$, and the results of this coding are placed in the columns E_{m+1} ... E_{2m}; i.e., the recoded E_1 is placed in E_{m+1}, etc.

C.2 Concatenating Columns

The Data ▶ Concatenate command combines two or more text columns into a single text column. For example, if C6 contains `m`, `m`, `m`, `f`, `f`, reading first to last entry, and C7 contains `to`, `ta`, `ti`, `to`, `ta`, then the entries in the Data ▶ Concatenate dialog box shown in Display C.2.1 result in a new text column C8 containing the entries `mto`, `mta`, `mti`, `fto`, `fta`.



Display C.2.1: Dialog box for concatenating text columns.

In the session environment, the **concatenate** command is available for this operation. The general syntax of the **concatenate** command is

concatenate E_1 ... E_m in E_{m+1}

where E_1, ..., E_m, are text columns, and E_{m+1} is the target text column.

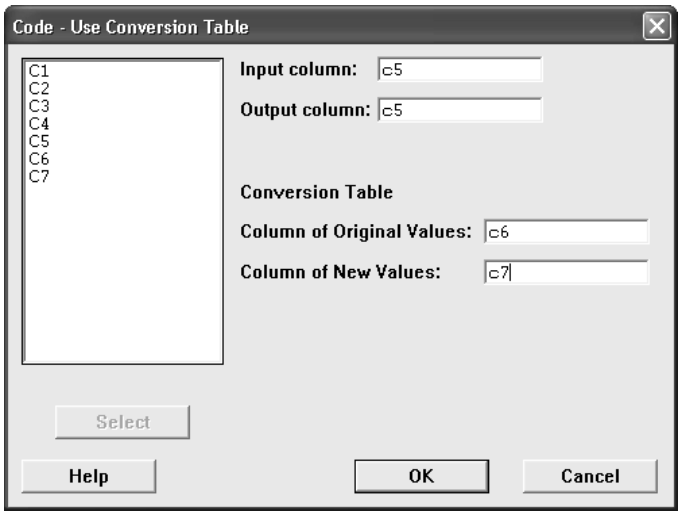
C.3 Converting Data Types

The Data ▶ Code ▶ Use Conversion Table command is used to change text data into numeric data and vice versa. As dealing with text data is a bit more difficult in Minitab, we recommend either converting text data to numeric before input or using this command after input to do this.

For example, in the worksheet `marks` (Display I.4) suppose we want to change the gender variable from text, with male and female denoted by `m` and `f`, respectively, to a numerical variable with male denoted by 0 and female by 1. To do this, we must first set up a *conversion table*. The conversion table comprises two columns in the worksheet, where one column is text and contains the text values used in the text column, and the second column is numeric and contains the numerical values that you want these changed into. For example, suppose we have entered columns C6 and C7 in the `marks` worksheet, as shown in Display C.3.1. The Data ▶ Code ▶ Use Conversion Table command produces the dialog box shown in Display C.3.2, where we have indicated that we want to convert the text column C5 into a numeric column and that each `m` should become a 0 and each `f` should become a 1.

C6-T	C7
m	0
f	1

Display C.3.1: Columns C6 and C7 in the marks worksheet as a conversion table.



Display C.3.2: Dialog box for converting text column C5 of the marks worksheet into a numeric column with the conversion table given in columns C6 and C7.

The general syntax for the corresponding session command **convert** is

convert E_1 E_2 E_3 E_4

where E_1, E_2 are the columns containing the conversion table, E_3 is the column to be converted, and E_4 is the column containing the converted column.

C.4 History

Minitab keeps a record of the commands you have used and the data you have input in a session. This information can be obtained in the History folder available via the Project Manager Toolbar as in Display C.4.1, and which is available on the taskbar at the top of the Minitab window. Placing the cursor over each icon in this toolbar indicates that the Fourth icon corresponds to the Show History folder.

Display C.4.1. Project Manager Toolbar.

The commands can be copied from wherever they are listed and pasted into the Session window to be executed again, so that a number of commands can be executed at once without retyping. These commands can be edited before being executed again. This is very helpful when you have implemented a long sequence of commands and realize that you made an error early on. Note that even if you use the menu commands, a record is kept only of the corresponding session commands.

The **journal** command is available in the Session window if you want to keep a record of the commands in an external file. For example, entering

```
MTB >journal 'comm1'
Collecting keyboard input(commands and data)in file:
                                                        comm1.MTJ
MTB >read c1 c2 c3
DATA>1 2 3
DATA>end
 1 rows read.
MTB >nojournal
```

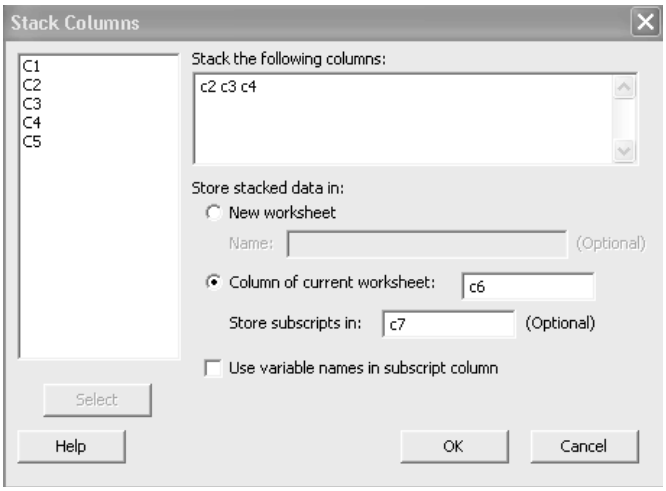
puts

```
read c1 c2 c3
1 2 3
end
nojournal
```

into the file `comm1.mtj`. The history is turned off as soon as the **nojournal** command is typed.

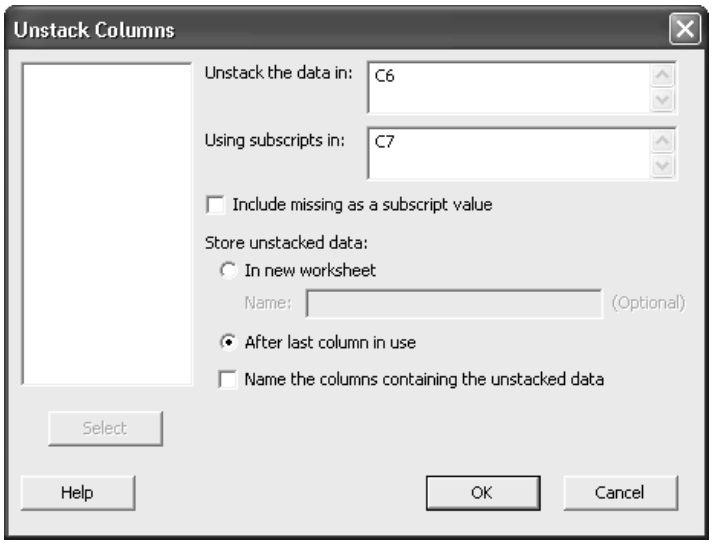
C.5 Stacking and Unstacking Columns

The Data ▶ Stack command is used to literally stack columns one on top of the other. For example, in the `marks` worksheet (Display I.4) the Data ▶ Stack ▶ Stack Columns command brings up the dialog box shown in Display C.5.1, which has been filled in to stack columns C2, C3, and C4 into C6 with the values in C2 first, followed by the values in C3 and then the values in C4. In C7, we have stored an index which indicates the column each value in C6 came from with a 1 every time a value came from C2, a 2 every time a value came from C3, and a 3 every time a value came from C4. It is not necessary to create such an index. Note if we check the Use variable names in subscript box, then instead of 1 in column C7 the text value C2 will appear, instead of 2 the text value C2 will appear, etc.



Display C.5.1: Dialog box for stacking columns.

To unstack values in a column by the values in an index column we use the Data ▶ Unstack command. For example, given the columns C6 and C7 of the `marks` worksheet as described above, the dialog box shown in Display C.5.2 unstacks C6 into three columns by the values in C7. The three columns are C8, C9, and C10. Note that they are identical to columns C2, C3, and C4, respectively. We must always specify a column containing the subscripts when unstacking a column.



Display C.5.2: Dialog box for unstacking columns.

In the Session window, the same results can be obtained using the **stack** and **unstack** commands. The general syntax for the **stack** command is given by

stack $E_1E_2 \ldots E_m$ into E_{m+1}

where E_1, E_2, ..., E_m denote the columns or constants to be stacked one on top of the other, starting with E_1, and with the result placed in column E_{m+1}. If we want to keep an index of where the values came from, then use the subcommand

subscripts E_{m+2}

which results in index values being stored in column E_{m+2}. The general syntax for the corresponding session command **unstack** is

unstack E_1 into $E_2 \ldots E_m$;
subscripts E_{m+1}.

where E_1 is the column to be unstacked, E_2, ..., E_m are the columns and constants to contain the unstacked column, and E_{m+1} gives the subscripts 1, 2, ... that indicate how E_1 is to be unstacked. Note that it is also possible to simultaneously unstack blocks of columns. We refer the reader to **help** or Help for information on this.

Index

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

$$z = \frac{\overline{x} - 450}{100/\sqrt{500}}$$

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 $\text{SE}_{b_1} = 6.31$

(b) $n = 25,\ \hat{y} = 13.0 + 6.10x,$ and $\text{SE}_{b_1} = 6.31$

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	0.00074
SAT Verbal	0.00164	0.00076
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	620	469	485	660	588	675
	517	649	209	404	738	628	609
	617	704	558	653	548		
5 months	490	395	402	177	475	617	616
	587	528	518	370	431	518	639
	368	538	519	506			
6 months	585	647	477	445	485	703	538
	465						

(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 CO2 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	203	229	208	146	230	215	153	205
	80	208	89	49	93	46	34	39	88
Medium	191	219	205	57	105	109	82	88	39
	94	206	197	68	44	203	139	72	67

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.